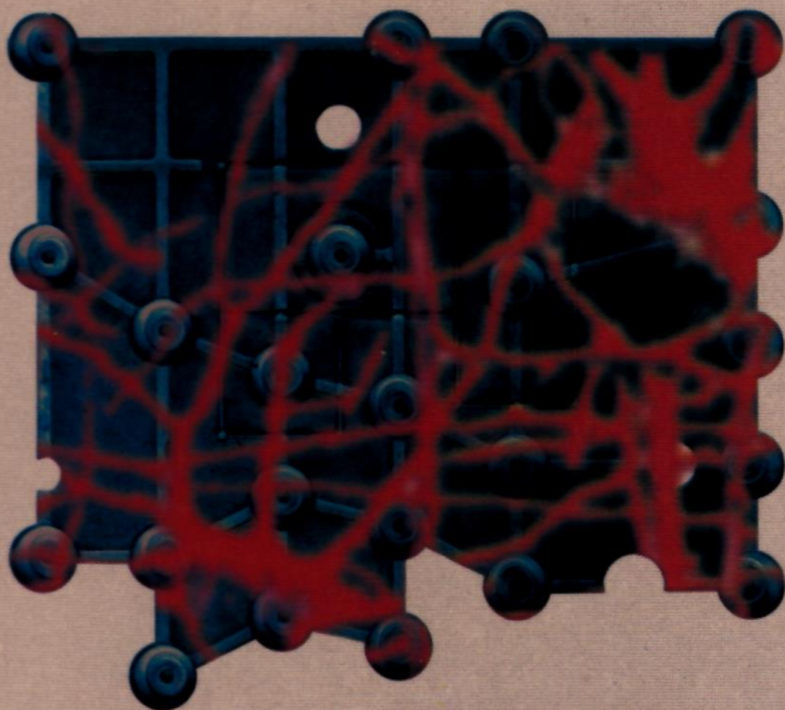
Ciencias cognitivas

Stephen R. Graubard *(comp.)*

El nuevo debate sobre la inteligencia artificial

Sistemas simbólicos y redes neuronales

Con contribuciones de
Hubert L. Dreyfus, Daniel C. Dennett,
Hilary Putnam y otros



gedisa
editorial

Stephen R. Graubard *(comp.)*

EL NUEVO DEBATE SOBRE LA INTELIGENCIA ARTIFICIAL

Serie CLA • DE • MA
Ciencias Cognitivas

**Editorial Gedisa ofrece
los siguientes títulos sobre**

CIENCIAS COGNITIVAS

STEPHEN R. GRAUBARD (COMP)	***El nuevo debate sobre la inteligencia artificial***
PAUL M. CHURCHLAND	***Materia y conciencia. Introducción contemporánea a la filosofía de la mente***
P. ENGEL (COMP)	***Psicología ordinaria y ciencias cognitivas***
F. VARELA, E. THOMPSON Y E. ROSCH	***De cuerpo presente***
DANIEL C. DENNETT	***La libertad de acción***
HILARY PUTNAM	***Representación y realidad***
DANIEL C. DENNETT	***La actitud intencional***
FRANCISCO J. VARELA	***Conocer***
	Realidad mental y mundos posibles

EL NUEVO DEBATE SOBRE LA INTELIGENCIA ARTIFICIAL

Sistemas simbólicos y redes neuronales

por

Stephen R. Graubard
(comp.)

gedisa
editorial

Título del original en inglés:
The artificial intelligence debate False starts, real foundations.

Publicado por MIT Press edition, Cambridge,
Massachusetts, London, England

Traducción: Carlos Reynoso
Diseño de cubierta: Sebastián Puiggrós

Segunda edición, febrero de 1999, Barcelona

Muntaner, 460, entlo., 1.ª
Tel. 93 201 60 00
08006 - Barcelona, España
e-mail: gedisa@gedisa.com
*http://*www.gedisa.com

ISBN: 84-7432-466-1
Depósito legal: B-6.397/1999

Impreso en Limpergraf
c/ Mogoda, 29-31. 08210 Barberà del Vallès

Impreso en España
Printed in Spain

Indice

Nota del traductor

En esta edición se ha optado por la terminología informática corriente que rige en la comunicación científica y en la actividad académica en informática en lengua castellana, antes que por los neologismos artificiosos que prevalecen en traducciones que no denotan familiaridad con la temática. De esta manera, se han conservado los términos ingleses que son de uso corriente y que no poseen una traducción aceptable y consensuada (*hardware*, *software*, etc.) y los vocablos para conectores lógicos (*and*, *or*, *else or*) que son simultáneamente cláusulas estándar de lenguajes de programación.

Lic. Carlos Reynoso
CONICET - Universidad de Buenos Aires

1

¿Una sola IA o muchas?

Seymour Papert

¿Hay una sola IA o más bien hay muchas? Ha habido un cambio espectacular en el tono de las discusiones sobre la inteligencia artificial; y ese cambio ha hecho que se perciba de repente con más claridad la existencia de modos de pensar divergentes en lo que alguna vez se presentó como un campo unificado. Los lectores de esta edición de *Dædalus* que no hayan estado en contacto con los desarrollos recientes pueden llegar a sorprenderse al ver cuántos de sus autores han escogido concentrarse en las divergencias que atraviesan el campo, y particularmente en una tendencia de la IA que ha venido a conocerse como conexionismo. No estarán solos en esta sorpresa. A fines de 1985 participé en un encuentro de planificación para coordinar un número de *Dædalus* sobre IA. En esa época yo sabía (y presumo que lo sabían también casi todos los que participaron del encuentro) que la actividad de investigación sobre temas «conexionistas» estaba en aumento. Pero yo habría desconfiado si alguien en el encuentro hubiera sugerido (nadie lo hizo) que esos temas pronto desbordarían las revistas técnicas, llegando hasta publicaciones tales como el *New York Times Book Review* —en la que el conexionismo se caracteriza como la contrarrevolución cognitiva[1]— y que se convertiría en el tópico de conversación central donde quiera que se discuta de IA o de ciencia cognitiva. El contenido de este número de *Dædalus* refleja este movimiento más de lo que lo hubiera hecho un plan deliberado: ha sucedido

---

Seymour Papert. Profesor de Tecnología de Medios y director del Grupo de Aprendizaje y Epistemología del Laboratorio de Medios del MIT.

algo intrigante y dramático, en una escala que excede la de la planificación de una revista. De este modo, cuando Stephen Graubard me invitó a contribuir con un artículo propio, no pude resistirme a utilizar el barullo conexionista como ocasión para discutir algunas cuestiones más amplias sobre la naturaleza de la inteligencia artificial y sobre la fascinación que ejerce sobre gente más interesada en la mente humana que en construir robots.

El campo de la inteligencia artificial se divide actualmente en lo que parecen ser varios paradigmas en contienda. Los contendientes actuales difieren en lo que respecta a las formas que se necesitan para capturar todas las formas de la inteligencia. Se encuentran abocados a la búsqueda de mecanismos de aplicación universal. Allen Newell, decano del procesamiento de la información, cree que él está muy cerca, que todo el conocimiento se puede formular como las reglas que están detrás de un tipo especial de programa conocido como «sistema de producción». Los autores del manifiesto conexionista contemporáneo, *Parallel Distributed Processing*,[2] no piensan estar tan cerca, pero expresan con confianza que el camino escogido (que se funda no en programas sino en entidades en red similares a las neuronas) proporcionará esos mecanismos universales.

No preveo el futuro como una victoria definitiva de alguno de los contendientes actuales. Lo que preveo es un cambio de marco que nos apartará de la búsqueda de mecanismos universales. Creo que tenemos mucho más que aprender del estudio de las diferencias entre las clases de conocimiento que de su similitud. Y sólo porque el conocimiento tenga lugar en cerebros no es una razón para argumentar, como lo hacen tanto los conexionistas como los programadores, que hay un mecanismo privilegiado y universal en algún nivel psicológicamente relevante.

Una analogía dramatiza lo que quiero decir con psicológicamente relevante. Un biólogo evolucionista puede intentar comprender cómo es que los tigres llegan a tener rayas. Y un biólogo molecular puede intentar comprender la forma en que se origina la vida en algún caldo primordial. Pero la forma en que la vida se origina no nos da información sobre el aspecto de un tigre. Sin embargo, esta falacia atraviesa el discurso intelectual de los conexionistas y los programadores. Los conexionistas hablan de experimentos a nivel de grupos pequeños de neuronas simuladas; y luego, casi simultáneamente, hablan sobre cómo es que se

puede caminar y pensar al mismo tiempo. Se presupone aquí que el multiprocesamiento es en ambos casos la misma clase de empresa. Los expertos en procesamiento de la información muestran sistemas de reglas que coinciden con el comportamiento de las personas y de las computadoras que resuelven problemas lógicos, y saltan de eso a afirmaciones como la de Allen Newell: «La psicología ha llegado a la posibilidad de una teoría unificada del conocimiento».

En ambos bandos se incurre en el mismo equívoco: el error categorial de suponer que la existencia de un mecanismo común proporciona tanto una explicación como una unificación de todos los sistemas, por complejos que sean, en los que ese mecanismo desempeñe un papel central. La tesis que sostengo aquí es que la IA necesita definirse de un modo que no la ponga en el peligro de cometer este error categorial. A medida que madure, veo que la inteligencia artificial desarrollará los marcos conceptuales que nos permitirán obtener una comprensión rigurosa no sólo de lo que es igual en actividades tales como enamorarse y jugar al ajedrez, sino de lo que es diferente entre ellas. La inteligencia artificial se convertirá en la metodología para pensar sobre formas de conocer.

En este ensayo utilizo un incidente en el desarrollo del conexionismo para ilustrar la actual resistencia de este campo ante esta forma de pensar sobre su identidad intelectual.

I

No llego a la discusión sobre el conexionismo como un observador neutral. De hecho, la versión estándar de esa historia me asigna un papel en una historia romántica cuyas resonancias con los cuentos de hadas seguramente contribuye, al menos un poco, al aura de excitación del conexionismo.

Había una vez dos ciencias hermanas de la nueva ciencia de la cibernética. Una hermana era natural y tenía rasgos heredados del estudio del cerebro, de la forma en que la naturaleza hace las cosas. La otra era artificial, relacionada desde el comienzo con el uso de las computadoras. Cada una de las ciencias hermanas procuraba construir modelos de la inteligencia, pero a partir de materiales muy diferentes. La hermana natural construía modelos (llamados redes neuronales) con neuronas matemáticamente

purificadas. La hermana artificial construía sus modelos con programas de computación.

En el primer florecer de su juventud ambas eran igualmente exitosas e igualmente codiciadas por pretendientes de otros campos del conocimiento. Juntas se llevaban bien. Pero sus relaciones cambiaron a principios de la década de 1960, cuando apareció un nuevo monarca, uno que poseía los cofres más grandes que se hubieran visto en el reino de las ciencias: el señor DARPA, la Agencia de Proyectos de Investigación Avanzados del Departamento de Defensa. La hermana artificial se tornó celosa y tomó la decisión de guardar para sí el acceso a los fondos de investigación del señor DARPA. La hermana natural debía ser eliminada.

El trabajo sangriento fue intentado por dos firmes seguidores de la hermana artificial, Marvin Minsky y Seymour Papert, puestos en el papel del cazador enviado a matar a Blancanieves y a traer su corazón como prueba. El arma no era una daga sino una pluma, mucho más poderosa, de la que salió un libro (*Perceptrons*[3]) destinado a probar que las redes neuronales nunca podrían cumplir su promesa de construir modelos de la mente: *sólo los programas de computación podrían hacerlo.* La victoria de la hermana artificial parecía asegurada. Y ciertamente, durante la siguiente década todas las recompensas del reino fueron para su progenie, de la cual la familia de los sistemas expertos hizo la mejor fama y fortuna.

Pero Blancanieves no estaba muerta. Lo que Minsky y Papert habían mostrado al mundo como prueba de su muerte no era su corazón; era el corazón de un cerdo. Para ser más literales: en el libro podía leerse la prueba de que la estrategia de las redes neuronales para construir modelos de la mente estaba muerta. Pero una mirada más atenta revela que ellos habían demostrado mucho menos que eso. El libro, por cierto, señalaba limitaciones muy serias de una cierta clase de redes (hoy en día conocidas como perceptrones de una sola capa); pero se encontraba equivocado en lo concerniente a la suposición de que esta clase de red era el corazón del conexionismo. *Parallel Distributed Processing*, admitiendo que la suposición de marras pueda haber sido un error honesto, incurre en un tono de cuentos de hadas al hablar sobre cómo eran las cosas «en los días de Minsky y Papert». En aquel tiempo y lugar tan alejados, todavía estaban por hacerse los

descubrimientos técnicos que abrirían la visión (el mito sustentador del modelo conexionista) de redes neuronales mucho más poderosas de lo que entonces podía imaginarse.

Las escrituras conexionistas presentan la historia con un final feliz. La hermana natural había sido cuidada en los laboratorios de unos pocos investigadores ardientes que sostuvieron su fe, incluso cuando el mundo en general estaba convencido de que la empresa era fútil. Quién (o qué) debe ser puesto en el papel del Príncipe Encantador es un problema que abordaré más adelante: ¿Quiénes son los protagonistas del *affaire* amoroso del conexionismo actual? ¿Quién despertó al conexionismo? ¿Y por qué ahora? ¿Qué pasará después? Pero por el momento es suficiente tomar nota de que el príncipe ha surgido de una relativa miseria y de la oscuridad para ganar la admiración de todos, excepto de unos pocos irritados partidarios de su hermana.

II

La narración parece hacer un requerimiento de culpabilidad o de inocencia: ¿Intentamos Minsky y yo matar al conexionismo? ¿Cómo nos sentimos ahora con su resurrección? Se necesita algo más complicado que un requerimiento así. Sí, ha habido *alguna* hostilidad en la energía puesta en la investigación reportada en *Perceptrons*, y hay *algún* grado de irritación ante la forma en que se ha desarrollado el nuevo movimiento; parte de nuestro impulso venía, como lo reconocimos derechamente en nuestro libro, del hecho de que la financiación y la energía de investigación se habían malgastado en lo que aún me sigue pareciendo (dado que la historia de mecanismos de red nuevos y poderosos es gravemente exagerada) un intento equivocado de métodos conexionistas en aplicaciones prácticas. Pero gran parte de la motivación de *Perceptrons* proviene de preocupaciones más fundamentales, muchas de las cuales atraviesan la frontera entre los diseñadores de redes y los programadores.

Una de estas preocupaciones tiene que ver con el hecho de que hay que hallar un equilibrio apropiado entre romanticismo y rigor en la búsqueda de la inteligencia artificial. Muchas empresas serias nunca despegarán del suelo si los pioneros se limitan a discutir en público sólo aquello que pueden demostrar rigurosa-

mente. Piénsese, por ejemplo, en el desarrollo de las máquinas voladoras. La excitación suscitada cuando los hermanos Wright hicieron su primer vuelo tuvo un gran elemento romántico. Y está bien que haya sido así: es difícil respetar a esos críticos que se lamentaron de que un corto salto en una playa no probaba la viabilidad de una transportación aérea útil. Cuando el éxito final no se puede tomar como criterio para juzgar los pasos iniciales, el problema de desarrollar una metodología crítica sensible es una parte esencial y a menudo delicada de toda empresa que se salga de lo ordinario.

En el caso de la inteligencia artificial, el problema de la evaluación crítica de resultados parciales se complica por el hecho de que una inteligencia pequeña no se reconoce con facilidad como inteligencia. En inglés incluso tenemos una palabra especial para ella: aunque un vuelo corto todavía se cuenta como un vuelo, una inteligencia pequeña todavía se cuenta como estupidez, y en las etapas iniciales de la IA (en las que todavía se encuentra) esto es todo lo que puede esperarse. ¿Cómo entonces decide uno si la última «estupidez» de una máquina se debe contar como un paso hacia la inteligencia? La metodología que Minsky y yo usamos en *Perceptrons* se explica mejor mediante un ejemplo.

Parallel Distributed Processing reporta un experimento en el cual una máquina simulada (la llamaré Exor)* aprendía a decir si dos datos ingresados, cada uno de los cuales podía ser un cero o un uno, eran diferentes. El proceso de aprendizaje de Exor consumía 2.232 repeticiones de un ciclo de entrenamiento; en cada repetición se presentaba a la máquina con una de cuatro posibles combinaciones de entradas (uno-uno, cero-cero, uno-cero, cero-uno) y una señal de retroalimentación para indicar si ella ha dado la respuesta correcta («no» para la primera y «sí» para las otras). ¿Sagaz o estúpida? ¿Debe estar uno más impresionado por el hecho de que la cosa «aprendiera» a fin de cuentas, o porque lo hacía tan lenta y laboriosamente?

Hubo un tiempo, en los días iniciales de la cibernética, en el que una máquina que no hiciera nada parecido al aprendizaje podía impresionar. Hoy en día se necesita algo más para ser significativo, y en este caso ese algo más tiene una relación muy estrecha con

*XOR, pronunciado como si se escribiera exor, es una abreviatura computacional de «o exclusivo» (es decir, «esto o aquel pero no ambos»). Esto lo convierte en el nombre perfecto para nuestra máquina simulada.

nuestra alegoría. Exor es una red neuronal, y sucede que el trabajo que aprende a ejecutar, con toda su simplicidad, es una de esas cosas que una red de una sola capa no puede hacer. Saber esto convierte al dilema de juzgar a Exor en el dilema mayor de evaluar al conexionismo. Si usted quiere creerlo, Exor le permite proclamar que «Blancanieves vive». Si no, el lento ritmo de aprendizaje de Exor le permite suspirar «pero a duras penas». *Perceptrons* expone un plan de acción muy diferente: en lugar de preguntar si las redes son buenas, preguntamos *para qué son buenas*. El foco de interés pasa de las generalidades sobre clases de máquinas a cuestiones específicas sobre clases de tareas. Desde este punto de vista, Exor suscita preguntas como ¿Qué tareas se pueden aprender más rápido y cuáles se pueden aprender aun más lentamente en esta máquina? ¿Podemos desarrollar una teoría de las tareas que explique por qué se necesitan 2.232 repeticiones en este acto de aprendizaje en particular? El cambio en la perspectiva es abrupto: el interés se ha trasladado desde los juicios sobre la máquina hacia el uso del rendimiento de la máquina en tareas particulares como un modo de aprender más sobre la naturaleza de las tareas. Esta traslación se refleja en el subtítulo de nuestro libro, *Perceptrons: An Introduction to Computational Geometry*. Encaramos nuestro estudio de las redes neuronales observando cuidadosamente las clases de tareas en función de las cuales se ha promovido su uso en su momento. Dado que la mayoría de ellas se situaban en el área de reconocimiento visual de patrones, nuestra metodología nos condujo a la construcción de teorías sobre esos patrones. Para nuestra sorpresa, nos encontramos de pronto trabajando en una nueva área de problemas para la investigación geométrica, preocupada en comprender por qué un mecanismo de reconocimiento determinado podía ejecutar con facilidad algunas tareas de reconocimiento, mientras que otras computaciones resultaban extremadamente costosas en términos del número de repeticiones o del monto de maquinaria requeridos para una tarea. Por ejemplo, un perceptrón pequeño de una sola capa puede distinguir con facilidad entre triángulos y cuadrados, pero se necesita una red muy grande para aprender si lo que se le pone delante es un solo objeto conectado o algo compuesto por muchas partes.

Nuestra sorpresa al encontrarnos trabajando en geometría fue una sorpresa agradable. Reforzó nuestra sensación de estar inau-

gurando un campo nuevo, en vez de estar clausurando un campo viejo. Pero aunque el paso de juzgar los perceptrones en forma abstracta a juzgar las tareas que ejecutaban pueda parecer puro sentido común, tomó un largo tiempo. Tan largo, de hecho, que estamos ahora un poco sorprendidos al observar la resistencia que los conexionistas actuales oponen a reconocer la naturaleza de nuestro trabajo, y la naturaleza del campo de problemas al que sus propias investigaciones puedan conducir eventualmente.

III

La inteligencia artificial, como cualquier otra empresa científica, ha construido una cultura científica. El modo de trabajo que usamos en *Perceptrons* corre contra la corriente de esa cultura, en cuyo desarrollo participamos nosotros mismos.

La búsqueda de la universalidad de los mecanismos se oscurece como rasgo penetrante de la cultura de la IA debido a la circunstancia de que todas las demostraciones exitosas, hechas tanto por programadores como por conexionistas, ejecutan tareas harto específicas en dominios harto estrechos. Por cierto, los teóricos de la IA proclaman a veces como un descubrimiento importante la teoría de que la especificidad del dominio no es una limitación de las máquinas sino una característica de la inteligencia. Sin embargo, la energía teorética de la IA no se ha encaminado a comprender las diferencias entre dominios específicos, sino a encontrar formas generales para los contenidos específicos.

Un rasgo universalista gana robustez si tiene múltiples raíces. Entre las raíces más profundas ha de estar la naturaleza mítica de la empresa original de la IA: la mente construyendo mentes. El deseo de universalidad ha sido alimentado asimismo por el legado de los científicos que construyeron la IA, en su mayor parte matemáticos. Y fue nutrido además por las circunstancias materiales más cotidianas de la financiación. Hacia 1969, la fecha de publicación de *Perceptrons*, la IA no operaba en el vacío de una torre de marfil. El dinero estaba en el candelero. Y mientras este hecho empujaba el campo hacia una preferencia por los logros a corto plazo, también premiaba afirmaciones en el sentido de que las inversiones de los financiadores darían frutos más allá de los productos inmediatos.

Su universalismo hizo casi inevitable que la IA se apropiara de nuestro trabajo como si se tratara de una prueba de que las redes neuronales eran universalmente malas. No pensábamos que nuestra misión fuera matar a Blancanieves; lo veíamos más bien como una forma de comprenderla. De hecho, más de la mitad de nuestro libro está dedicada a hallazgos «properceptrón» acerca de cosas muy sorprendentes y hasta entonces desconocidas que los perceptrones pueden hacer. Pero en una cultura configurada para juicios universales sobre los mecanismos, ser comprendido puede ser un destino tan malo como la muerte. Una comprensión verdadera de lo que los mecanismos pueden hacer acarrea demasiadas implicaciones sobre lo que no pueden hacer.

El mismo rasgo de universalismo lleva a las nuevas generaciones de conexionistas a contemplar sus propios experimentos de micronivel, tales como Exor, como una pantalla proyectiva en la cual observar las macrocuestiones más importantes de la filosofía de la mente. El error categorial de buscar explicaciones de las rayas del tigre en la estructura del ADN no es un error aislado. Está sólidamente enraizado en la cultura de la IA.

IV

La inmodestia de usar el cuento de Blancanieves como metáfora me ha permitido hablar de la contrarrevolución conexionista sin decir exactamente lo que el conexionismo es o contra qué revoluciona. Se necesita un poco más de detalle técnico para situar el conexionismo en el campo mayor de las ciencias de la mente.

La tarea concreta de reconocer la igualdad de dos insumos binarios sería trivial para un programador. El primero de los rasgos notables detentados por Exor es que nadie la programó; fue «entrenada» para hacer esa tarea mediante un proceso estrictamente conductista de asociación externa de estímulos con refuerzos. Podría haber sido entrenada por alguien que estuviera de acuerdo con las invectivas de Watson en contra de pensar en las interioridades de un sistema. Pero si éste fuera su único mérito como modelo de procesos mentales, el gran número de repeticiones negaría su interés: máquinas diseñadas específicamente para simular reflejos condicionados lo han logrado tras un número de repeticiones psicológicamente más plausible.

La pretensión de universalidad de Exor es un rasgo más fuerte. Exor es pequeña y tiene poca potencia, pero otorga sustento a la visión de máquinas más grandes que están construidas conforme al mismo principio y que aprenderán cualquier cosa que sea susceptible de aprenderse sin ninguna disposición innata para adquirir conductas particulares. La perspectiva de esos comportamientos deviene una reivindicación de algo más que de las redes neuronales. Promete una reivindicación del conductismo en contra de Jean Piaget, de Noam Chomsky y de todos los estudiosos de la mente que criticaron el universalismo inherente a la tabula rasa conductista. El conductismo había sido vencido en otra versión de la historia de Blancanieves, pero la respuesta de la psicología académica al conexionismo puede llegar a ser un ejemplo clásico del retorno de lo reprimido.

El conexionismo hace más que traer de vuelta a un conductismo pasado de moda. Lo trae de vuelta en una forma que ofrece una reconciliación con el pensamiento biológico acerca del cerebro. La estructura de la máquina refleja, aunque en una forma abstracta, un cierto modelo de cómo podría estar concebiblemente constituido el cerebro por neuronas. Aunque los experimentos concretos de Exor, de hecho, son ejecutados por programas de computadora, esos programas pretenden representar qué pasaría si uno conectara redes de unidades que se suponen similares a las neuronas en el siguiente sentido. Cada unidad de la red recibe señales de las otras o de unidades sensoriales conectadas al mundo exterior. En un tiempo dado, cada unidad posee cierto nivel de activación que depende del peso sumado de los estados de activación de las unidades que le envían información, y las señales que se envían a lo largo del «axón» de las unidades reflejan su estado de activación. El aprendizaje tiene lugar mediante un proceso que ajusta los pesos (fuerzas de conexión) entre las unidades; cuando los pesos son diferentes, los patrones de activación producidos por un insumo dado serán diferentes, y, por último, la salida (respuesta) ante un insumo (estímulo) cambiará. Este rasgo otorga a las máquinas de la familia de Exor un sabor biológico que seduce fuertemente al espíritu de nuestro tiempo, sin apartarse demasiado de la simplicidad conductista: aunque uno se tiene que referir a estructuras parecidas a las neuronas para construir la máquina, sólo se piensa en términos de estímulos, respuestas y una señal de retroalimentación para operarla.

V

Esta representación del conexionismo como conductismo en ropaje computacional ayuda a poner a *Perceptrons* en perspectiva: los problemas que discute son una forma moderna para un viejo debate acuñado originariamente como una discusión humanística y filosófica entre asociaciones y retomada después como una discusión sobre el conductismo. Tales debates giran a menudo en torno a aserciones de la forma «*Comenzando con nada excepto* (asociaciones, estímulo y respuesta, o lo que fuere), *nunca podrás lograr* (ideas generales, lenguaje, o lo que fuere)». Las discusiones de este tipo han sido más o menos atractivas, pero nunca han estado siquiera cerca de ser conducentes a estándares de rigor que parecieran normales a gente entrenada como matemáticos, como Minsky y yo lo somos. Además, ¿cómo podría siquiera formularse la discusión con un asomo de rigor en ausencia de una teoría firme del pensamiento humano? ¿Y cómo se podría avanzar hacia esa teoría firme sin saber si se pueden derivar ideas generales o lo que fuere de asociaciones o lo que fuere?

En su sentido más estrecho, la intención de *Perceptrons* ha sido evitar en el estudio del «pensamiento de máquina» algunas de las dificultades del tipo el-huevo-y-la-gallina que han atestado el pensamiento sobre el pensamiento humano. La estrategia consistía en estudiar cierta clase de máquinas computacionales que fueran lo suficientemente poderosas para capturar un fragmento significativo de los logros contemporáneos en IA, pero también lo suficientemente simples para hacer posible, con las pocas herramientas analíticas de que disponíamos, un análisis matemático riguroso de sus capacidades. Elegimos la clase de máquinas por las cuales llamamos el libro (en honor de Frank Rosenblatt): los perceptrones se definen en el libro como una clase especial y simple de red neuronal de la misma familia que Exor. Los perceptrones son demasiado simples para ser interesantes por derecho propio como modelos de procesos mentales. Pero el paso más promisorio hacia el desarrollo de herramientas lo suficientemente fuertes para analizar sistemas más complejos, incluyendo la mente humana, parecería presuponer la comprensión profunda de casos tan simples como los perceptrones. Muchos lectores, todos quizás exceptuando a los matemáticos, quedarán perplejos al saber qué simples pueden ser las máquinas y qué difícil es sin

embargo comprender plenamente sus capacidades. Me asusta pensar sobre lo duro que es confirmar o rechazar nuestras intuiciones a propósito de la capacidad de los perceptrones.

Tanto Minsky como yo conocemos los perceptrones extremadamente bien. Trabajamos sobre ellos muchos años antes que concibiéramos nuestro proyecto conjunto de comprender sus límites; más aún, nos encontramos por primera vez en una conferencia en la cual ambos presentamos artículos con un grado improbable de superposición de contenidos a propósito de lo que las máquinas similares a los perceptrones podían hacer. Con este trasfondo, debíamos estar en una posición excepcional para formular conjeturas sobre los perceptrones. Pero cuando nos desafiamos a probar nuestras intuiciones nos tomó años de lucha tomar cada una de ellas, probarla o descubrir que estaba seriamente equivocada.

Me ha quedado un profundo respeto por lo extraordinariamente difícil que es llegar a estar seguro de lo que un sistema computacional puede o no puede hacer. Me maravilla esa gente que parece tan segura en sus convicciones intuitivas, o con sus argumentos retóricos menos-que-rigurosos sobre computadoras, redes neuronales o mentes humanas. Un área en donde la intuición parece necesitar particularmente un análisis riguroso es en el tratamiento de la noción románticamente atractiva de proceso holístico.

VI

En la historia de la psicología, el conductismo y el holismo (o gestaltismo) se han considerado opuestos polares. El conductismo fragmenta la mente en una miríada de átomos separados de tamaño mucho menor de lo que permitiría el sentido común. El holismo y el gestaltismo insisten en que los átomos psicológicos son más grandes de lo que el sentido común piensa. De tal modo, es sumamente llamativo que el conexionismo tenga facetas que resultan atractivas para todas estas escuelas de pensamiento.

El título de la actual biblia del conexionismo, *Parallel Distributed Processing*, yuxtapone dos cualidades que para el movimiento conexionista son ciertamente características primarias de todo lo natural y quizá de las versiones artificiales efectivas de la inteligencia. *Parallel* se refiere a la cualidad de mantener activos varios

procesos al mismo tiempo: así como la gente camina y habla al mismo tiempo, muy probablemente desarrolle un número enorme de procesos mentales concurrentes, la mayoría inconscientes. *Distributed* se refiere a la cualidad de no estar localizado: en las computadoras convencionales, los elementos de información se almacenan en lugares particulares, netamente separados entre sí; en las redes neuronales, la información se desparrama (en principio, una nueva pieza de aprendizaje puede involucrar cambios en todas partes). Gran parte de la noción de que hay procesos profundos en el funcionamiento de las redes se relaciona con la idea de que lo que el discurso ordinario y la teoría cognitiva tradicional describen erróneamente como elementos atómicos de información, se halla representado en forma holística y es evocable de esa manera.

Lo paralelo más lo distribuido *se siente* bien. Pero el trabajo con los perceptrones nos ha tornado cautelosos ante formas en las que las dos cualidades, más que hallarse en dulce armonía, se encuentran en tensión. No es difícil cambiar las percepciones para hacer que la yuxtaposición se sienta intuitivamente problemática. En la vida ordinaria, la costumbre de separar las actividades en habitaciones y oficinas se funda en la experiencia de las consecuencias perturbadoras de dejar que todo pase donde quiera y al mismo tiempo. Pero el conexionismo se erige sobre la teoría (que Sherry Turkle llama un mito sustentador) de que una comprensión más profunda revelaría la ingenuidad de esas analogías cotidianas. Así como la física moderna nos enseña a no proyectar nuestro sentido de sus sucesos macroscópicos al mundo subatómico, así también la comprensión más profunda de las redes nos enseñará que nuestras metáforas de la organización macroscópica de las redes pueden ser igualmente equívocas.

Por cierto, se pueden hallar analogías en la ciencia física que van muy en contra de las intuiciones amorfas sobre la interferencia, sobre la forma en que los diferentes procesos se perturban entre sí. Las vibraciones de todas las ondas de radio y televisión pasan a través del mismo espacio al mismo tiempo y, sin embargo, los circuitos de sintonía pueden separarlas. Aún más incomprensible, si es que no francamente chocante para el sentido común, es el holograma, que registra una figura tridimensional en forma totalmente distribuida: si parte del registro holográfico se destruye, no se pierde una parte en particular de la figura, sino que hay una degradación uniforme de la calidad.

Estos ejemplos dicen lisa y llanamente que en el mundo físico hay precedentes de la superposición distribuida. En el universo hay tantas cosas holísticas que el concepto de red neuronal distribuida no se puede rechazar basándose en principios intuitivos generales. Pero no todo es holístico, y la opinión de sentido común (y aun la filosófica) es de poca utilidad para mapear qué es lo que lo es. Se necesita una investigación específica, a veces de una naturaleza matemática sutil y muy técnica, para determinar si la representación holística es posible en una situación específica y si (donde puede ser hecha) no hay un exorbitante precio que pagar. La máquina Exor ilustra, en un caso simple, el concepto de costo del holismo.

La tarea que aprendió Exor se puede ver como una superposición de dos aprendizajes en la misma red: aprendizaje para decir *sí* ante uno-cero y aprendizaje para decir *no* a cero-uno. Un hecho importante es que cada una de esas tareas, tomada separadamente, es mucho más fácil de aprender que la tarea combinada. Y este no es un fenómeno ocasional: Exor es un caso muy suave de costo de distribución contraído. Uno de los resultados de la investigación de *Perceptrons*, que requirió cierto empeño matemático, muestra que en ciertas situaciones el grado de dificultad de tareas superpuestas puede superar la dificultad de cada tarea separada por factores grandes y arbitrarios.

La instancia romántica sería hacer una nueva red que no fuera ya un perceptrón y presumir su inocencia hasta que se la pruebe culpable del peligro de costo de superposición. En suma, la literatura conexionista lo hace incluso cuando reporta experimentos en los que las nuevas redes muestran signos empíricos de costos tales como los que Exor manifiesta suavemente. La instancia rigurosa presume la posibilidad de la culpa hasta que se pueda establecer la inocencia: se considera que los teoremas probados sobre los perceptrones muestran qué clases de fenómenos deben evitarse antes de poder hacer aserciones con seguridad.

VII

Dije al principio que ofrecería algunas ideas sobre el Príncipe Encantador. ¿Quién despertó al conexionismo? ¿A qué se debe este brote de interés y actividad? ¿Por qué ahora? Utilizaré mis especulaciones sobre estos temas para abordar una importante cuestión: ¿Qué es lo que vendrá después?

Una reseña puramente técnica del despertar de Blancanieves reza como sigue: En los viejos días de Minsky y Papert, los modelos de redes neuronales se hallaban limitados sin esperanza por el encanijamiento de las computadoras disponibles en la época y por la falta de ideas sobre cómo hacer que una red que no fuera de las más simples aprendiera algo. Ahora las cosas han cambiado. Computadoras poderosas, masivamente paralelas, pueden implementar redes muy grandes, y nuevos algoritmos pueden hacer que aprendan. No se necesita un Príncipe Encantador para esta historia.

Yo no lo creo. Las demostraciones recientes más o menos influyentes de nuevas redes corren todas en computadoras pequeñas y pudieron haberse hecho con facilidad en 1970. Exor es un «problema de juguete» ejecutado para estudio y demostración, pero los ejemplos discutidos en la literatura siguen siendo muy pequeños. Ciertamente, Minsky y yo, en una discusión más técnica de la misma historia (añadida como capítulo nuevo a una reedición de *Perceptrons*), sugerimos que la estructura entera de las teorías conexionistas recientes podría estar construida sobre arena: se basa en problemas de tamaño de juguete sin un análisis teorético que muestre que la performance se mantendrá cuando los modelos posean una escala realista. Los autores conexionistas no alcanzan a leer nuestro trabajo como una advertencia en el sentido de que las redes, como los programas de «fuerza bruta» basados en procedimientos de búsqueda, escalan muy mal.

Se necesita una explicación más sociológica. Las supercomputadoras masivamente paralelas juegan un papel muy importante en la resurrección conexionista. Pero lo veo como un rol más cultural que técnico, como otro ejemplo de mito sustentador. El conexionismo no utiliza las nuevas computadoras como máquinas físicas; deriva su fuerza de la «computadora en el cerebro», del conocimiento público, en gran parte no técnico, sobre las supercomputadoras.

Contemplo la relación entre el conexionismo y la biología en parecidos términos. Aunque sus modelos utilizan metáforas biológicas, no dependen de los hallazgos técnicos de la biología más de lo que dependen de las supercomputadoras modernas. Pero aquí también hay un fenómeno poderoso y resonante. La biología es cada vez más el lugar de la mayor excitación. Y las neurociencias invaden el territorio de la psicología académica del

mismo modo en que la psicofarmacología invade el territorio de la psicología clínica.

También veo una resonancia cultural más sutil, pero no menos relevante. Hay un giro generalizado, que se aparta del racionalismo a ultranza propio de la época del primer eclipse conexionista, y una atracción resurgente hacia formas de pensar más holísticas. La discusión teorética actual en la literatura conexionista puede no estar conectada en un sentido estricto a tales tendencias de la moda intelectual. Pero nuevamente aquí, los conceptos del mito sustentador y de la resonancia cultural son pertinentes: esta vez, quizás, en un proceso en dos vías de soporte mutuo.

Voilà el Príncipe Encantador: un compuesto de tendencias culturales. Las resonancias reduccionistas de mi discusión no socavan mis buenos deseos de una unión feliz con Blancanieves. En la ciencia cognitiva, el nuevo sentimiento de excitación que está desplazando a un agotamiento letárgico asegurará la fertilidad de la unión. Pero el impacto del conexionismo vendrá menos de las ideas que engendra que de una atención despierta hacia los problemas que evita.

Notas

[1] James J. Greeno. «The Cognition Connection", *New York Times Book Review*, 4 de enero de 1987.

[2] David E. Rumelhart, James L. McClelland y el PDP Research Group, *Parallel Distributed Processing* (Cambridge: MIT Press, 1986).

[3] Marvin Minsky y Seymour Papert, *Perceptrons: An introduction to Computational Geometry* (Cambridge: MIT Press, 1969).

[4] Rumelhart, McClelland y el PDP Research Group, *Parallel Distributed Processing*, pág. 111.

2
Fabricar una mente versus modelar el cerebro: la inteligencia artificial se divide de nuevo

Hubert L. Dreyfus y Stuart E. Dreyfus

[N]ada me parece más posible que el hecho de que la gente llegará algún día a la opinión definitiva de que no hay copia en el ... sistema nervioso que corresponda a un pensamiento particular, *una idea o memoria* particular.[1]

Ludwig Wittgenstein (1948)

[La] información no está almacenada en particular en ninguna parte. Más bien, está almacenada por doquier. La información se piensa mejor como «evocada» que como «hallada».[2]

David Rumelhart y Donald Norman (1981)

A comienzos de la década de 1950, a medida que surgían las máquinas de calcular, unos pocos pensadores pioneros comenzaron a darse cuenta de que las computadoras personales podían ser más que masticadoras de números. En ese momento surgieron y lucharon por su reconocimiento dos visiones opuestas de lo que podían ser las computadoras, cada una con su programa

Hubert L. Dreyfus. Profesor de filosofía en la Universidad de California en Berkeley.
Stuart E. Dreyfus. Profesor de Ingeniería Industrial e Investigación Operativa en la Universidad de California en Berkeley.

de investigación correlativo. Una facción veía las computadoras como sistemas para manipular símbolos mentales; la otra, como un medio para modelizar el cerebro. Una buscaba usar las computadoras para instanciar una representación formal del mundo; la otra, para simular las interacciones de las neuronas. Una tomó la resolución de problemas como su paradigma de la inteligencia; la otra, las estadísticas. Una escuela era la heredera de la tradición racionalista y reduccionista en filosofía; la otra se veía a sí misma como una neurociencia idealizada y holística.

El grito de convocatoria del primer grupo era que tanto las mentes como las computadoras digitales eran sistemas físicos de símbolos. Hacia 1955 Allen Newell y Herbert Simon, trabajando en la Rand Corporation, llegaron a la conclusión de que unas cadenas de bits manipulados por una computadora digital podían estar en lugar de cualquier otra cosa: números, desde luego, pero también rasgos del mundo real. Más aún, los programas se podían usar como reglas para representar relaciones entre esos símbolos, de modo que el sistema pudiera inferir nuevos hechos sobre los objetos representados y sus relaciones. Como lo ha dicho Newell recientemente en su reseña de la historia de la IA,

> El campo de la computadora digital definía las computadoras como máquinas que manipulaban números. La gran cosa era, decían los adherentes, que cualquier cosa puede codificarse mediante números, incluso instrucciones. En contraste, los científicos de la IA vieron las computadoras como máquinas que manipulaban símbolos. La gran cosa era, decían, que todo puede codificarse como símbolos, incluso los números.[3]

Esta forma de mirar la computadora devino la base de una forma de mirar las mentes. Newell y Simon lanzaron la hipótesis de que el cerebro humano y la computadora digital, aunque totalmente diferentes en estructura y mecanismos, poseían, a cierto nivel de abstracción, una descripción funcional común. A este nivel, tanto el cerebro humano como una computadora digital adecuadamente programada se podían considerar dos instancias diferentes de una sola clase de dispositivo: un dispositivo que generaba conducta inteligente manipulando símbolos por medio de reglas formales. Newell y Simon estipulaban esta perspectiva como una hipótesis:

Hipótesis del Sistema / / Físico de Símbolos: Un sistema físico de símbolos posee los medios necesarios y suficientes para una acción inteligente general.

Por «necesario» queremos decir que cualquier sistema que exhiba inteligencia general probará bajo análisis ser un sistema físico de símbolos. Por «suficiente» queremos decir que cualquier sistema físico de símbolos de tamaño suficiente se puede organizar para exhibir inteligencia general.[4]

Newell y Simon remontan las raíces de su hipótesis hasta Gottlob Frege, Bertrand Russell y Alfred North Whitehead;[5] pero Frege y compañía eran a su vez herederos de una larga tradición atomística y racionalista. Descartes ya había pensado que toda la comprensión consistía en la formación y manipulación de representaciones apropiadas, que esas representaciones podían analizarse en sus elementos primitivos (*naturas simplices*) y que todos los fenómenos podían comprenderse como combinaciones complejas de esos elementos simples. Más aún: en la misma época, Hobbes suponía implícitamente que los elementos eran componentes formales relacionados mediante operaciones puramente sintácticas, de modo que el razonamiento podía reducirse al cálculo. «Cuando un hombre razona, no hace más que concebir una suma total a partir de una adición de parcelas», escribía Hobbes, «pues la RAZON no es otra cosa que cálculo...».[6] Finalmente Leibniz, al elaborar la idea de la matesis —la formalización de todo— buscaba sustentar el desarrollo de un sistema universal de símbolos de modo tal que «podamos asignar a cada objeto un número característico determinado.»[7] De acuerdo con Leibniz, al analizar nosotros analizamos los conceptos en sus ele-mentos más simples. A fin de evitar un retorno a elementos más y más simples, debe haber simples últimos en términos de los cuales se puedan comprender todos los objetos complejos. Más aún, si se han de aplicar conceptos al mundo, debe haber rasgos simples que esos elementos representan. Leibniz vislumbró «una especie de alfabeto de pensamientos humanos»[8] cuyos «caracteres deben mostrar, cuando se utilizan en demostraciones, alguna especie de conexión, agrupamiento u orden que también se halla en los objetos.»[9]

Ludwig Wittgenstein, inspirándose en Frege y Russell, enunció en su *Tractatus Logico-Philosophicus* la forma pura de esta concep-

ción sintáctica y representativa de la relación entre la mente y la realidad. Definió el mundo como la totalidad de los hechos atómicos lógicamente independientes:

1.1 El mundo es la totalidad de los hechos, no de las cosas.

Los hechos, a su vez —sostenía— se podían analizar exhaustivamente en objetos primitivos.

2.01. Un hecho atómico es una combinación de objetos...

2.0124. Si todos los objetos están dados, luego por *ello* todos los hechos atómicos están dados.

Estos hechos, sus constituyentes y sus relaciones lógicas —argumentaba Wittgenstein— se representaban en la mente.

2.1. Hacemos para nosotros mismos retratos de las cosas.

2.15. El hecho de que los elementos del retrato se combinen entre sí de un modo definido, representa que las cosas se combinan así entre ellas.[10]

Se puede concebir la IA como un intento de encontrar los elementos primitivos y las relaciones lógicas en el sujeto (hombre o computadora) que refleja como un espejo los objetos primitivos y las relaciones que constituyen el mundo. La hipótesis del sistema de símbolos físicos de Newell y Simon, en efecto, torna la visión wittgensteiniana (que es en sí la culminación de la tradición filosófica racionalista clásica) en una afirmación empírica y basa en ella un programa de investigación.

La intuición opuesta, que nosotros expondremos como la creación de la inteligencia artificial mediante el modelado del cerebro más que de la representación simbólica y mental del mundo, se inspiró no en la filosofía, sino en lo que pronto habría de llamarse neurociencia. Fue inspirada directamente por la obra de D.O. Hebb, quien en 1949 sugirió que una masa de neuronas podía aprender si, cuando la neurona A y la neurona B fueran simultáneamente excitadas, su excitación reforzara la conexión entre ellas.[11]

Esta línea fue continuada por Frank Rosenblatt, quien pensó que, dado que la conducta inteligente basada en nuestra representación del mundo parecía difícil de formalizar, la IA debería en lugar de eso intentar automatizar los procedimientos mediante los

cuales una red de neuronas aprendía a discriminar patrones y a responder adecuadamente. Como decía Rosenblatt,

> La presunción implícita [del programa de investigación de la manipulación de símbolos] es que es relativamente fácil especificar la conducta que esperamos ejecute el sistema, y que el desafío entonces es diseñar un dispositivo o mecanismo que efectivamente desarrolle ese comportamiento. [E]s al mismo tiempo más fácil y más provechoso axiomatizar el sistema físico y luego investigar este sistema para determinar su conducta, que analizar la conducta y luego diseñar un sistema físico con técnicas de síntesis lógica.[12]

Otra forma de expresar la diferencia entre los dos programas de investigación es decir que quienes buscaban representaciones simbólicas perseguían una estructura formal que diera a la computadora la habilidad para resolver cierta clase de problemas o discriminar cierto tipo de patrones. Rosenblatt, por otro lado, pretendía construir un dispositivo físico o simular ese dispositivo en una computadora digital que pudiera entonces generar sus propias habilidades:

> Muchos de los modelos de los que hemos oído discutir se refieren a la pregunta de qué estructura debe poseer un sistema si éste ha de exhibir una propiedad X. Esta es esencialmente una pregunta sobre un sistema estático...
>
> Una forma alternativa de mirar este problema es: ¿qué clase de sistema puede *desarrollar* la propiedad X? Pienso que podemos mostrar en cierto número de casos interesantes que la segunda pregunta se puede resolver sin tener respuestas para la primera.[13]

Ambas estrategias tuvieron éxito inmediato y novedoso. Para 1956 Newell y Simon lograron programar una computadora utilizando representaciones simbólicas para resolver simples enigmas y probar teoremas del cálculo proposicional. Sobre la base de estos resultados tempranos e impresionantes, pareció como si la hipótesis del sistema físico de símbolos estuviera a punto de confirmarse, y Newell y Simon se hallaban comprensiblemente eufóricos. Simon anunció:

> No es mi propósito sorprenderlos o impactarlos. Pero la forma más simple en que puedo resumirlo es diciendo que ahora hay en el mundo

máquinas que piensan, aprenden y crean. Más aún, su habilidad para hacer estas cosas se va a incrementar rápidamente –en un futuro cercano– hasta que el rango de problemas que puedan manejar sea coextensivo al rango de problemas a los que se ha aplicado la mente humana.[14]

El y Newell explicaban:

[T]enemos ahora los elementos de una teoría de la resolución de problemas heurística (en contraste con la algorítmica); y podemos usar esta teoría tanto para comprender los procesos heurísticos humanos como para simular dichos procesos con computadoras digitales. La intuición, la comprensión y el aprendizaje no son más posesión exclusiva de los humanos: cualquier computadora grande de alta velocidad puede ser programada para también exhibirlos.[15]

Rosenblatt puso a trabajar estas ideas en un tipo de dispositivo que él llamó perceptrón.[16] En 1956 Rosenblatt pudo entrenar un perceptrón para clasificar ciertos tipos de patrones como patrones similares, separándolos de otros patrones diferentes. En 1959 estaba alborozado y sentía que su estrategia había sido vindicada:

Me parece claro que el perceptrón nos presenta un nuevo tipo de autómata de procesamiento de información: por primera vez, tenemos una máquina que es capaz de tener ideas originales. Como un análogo del cerebro biológico, el perceptrón o, más precisamente, la teoría de la separabilidad estadística, parece más próxima a satisfacer los requerimientos de una explicación estadística del sistema nervioso que ningún otro sistema propuesto con anterioridad. Como concepto, parecería que el perceptrón ha establecido, más allá de toda duda, la practicabilidad y el principio de los sistemas no humanos capaces de encarnar funciones cognitivas humanas. El futuro de los dispositivos de procesamiento de información que opera sobre principios estadísticos, más que lógicos, me parece ahora claramente establecido.[17]

A principios de la década de 1960 ambas estrategias parecían igual de promisorias y se hicieron por igual vulnerables al efectuar afirmaciones exageradas. Pero los resultados de la guerra interna entre los dos programas de investigación eran sorprendentemente asimétricos. En 1970 la investigación de simulación del cerebro,

que poseía su paradigma en el perceptrón, se redujo a unos pocos esfuerzos aislados y subfinanciados, mientras que los que proponían usar las computadoras digitales como manipuladoras de símbolos poseían el control indiscutido de los recursos, programas de estudio, revistas y simposios en lo que constituía un programa de investigación floreciente.

Reconstruir cómo se suscitó este cambio es complicado, debido al mito del destino manifiesto que genera todo programa de investigación en marcha. Por ejemplo, contemplaríamos a los vencedores como si el procesamiento simbólico de información hubiese ganado debido a que se hallaba en el camino correcto, mientras que la estrategia de la red neuronal o conexionista hubiese perdido porque sencillamente no funcionaba. Pero esta versión de la historia es una ilusión retrospectiva. Ambos programas de investigación poseían ideas dignas de explorarse y ambos enfrentaban problemas profundos y no reconocidos.

Cada postura tenía sus detractores, y lo que ellos decían era básicamente lo mismo: ambas estrategias han demostrado que pueden resolver ciertos problemas fáciles, pero no hay razón para creer que alguno de los dos grupos pueda extrapolar sus métodos a la complejidad del mundo real. Por cierto, había evidencias de que a medida que los problemas se tornaban más complejos, la computación requerida por ambas estrategias crecía exponencialmente y pronto devenía intratable. En 1969 Marvin Minsky y Seymour Papert decían de los perceptrones de Rosenblatt:

> Los esquemas de Rosenblatt pronto se afianzaron, y muy pronto había tal vez cien grupos, grandes y pequeños, experimentando con el modelo.
>
> Los resultados de estos cientos de proyectos y experimentos eran en general frustrantes y sus explicaciones inconcluyentes. Habitualmente las máquinas trabajaban muy bien en problemas muy simples pero se deterioraban muy rápido a medida que las tareas que se les asignaban se volvían más duras.[18]

Tres años después, Sir James Lighthill, tras comentar el trabajo de los programas heurísticos como los de Simon y Minsky, llegó a una conclusión negativa sorprendentemente parecida:

> La mayoría de quienes trabajan en IA y en campos relacionados confiesan un acentuado sentimiento de frustración en lo que se ha

logrado en los últimos 25 años. Los estudiosos ingresaron al campo hacia 1950, y aun hacia 1960, con grandes esperanzas que están muy lejos de haberse realizado en 1972. En ningún área de ese campo los descubrimientos hechos hasta ahora produjeron el gran impacto que se había prometido...

[H]ay una causa general para las frustraciones que se han experimentado: no se han reconocido las implicancias de la «explosión combinatoria». Este es un obstáculo general que impide la construcción de un sistema sobre una base de conocimientos grande, que resulta del crecimiento explosivo de cualquier expresión combinatoria, correlativa a otras tantas formas de agrupar los elementos de la base de conocimiento de acuerdo con reglas particulares, a medida que el tamaño de la base se incrementa.[19]

Tal como lo han expuesto sucintamente David Rumelhart y David Zipser, «La explosión combinatoria te atrapa tarde o temprano, aunque a veces en forma diferente en paralelo que en modo serial».[20] Ambos bandos, tal como lo expresó James Fodor, habían caído en un juego de ajedrez tridimensional, pensando que se trataba de ta-te-ti. ¿Por qué entonces, apenas iniciado el juego, siendo tan poco lo que se conocía y habiendo tanto por aprender, un grupo de investigadores triunfó a expensas del otro? ¿Por qué, en esta encrucijada crucial, el proyecto de la representación simbólica se convirtió en el único juego en todo el pueblo?

Todos los que conozcan la historia del campo podrán señalar la causa proximal. Hacia 1965, Minsky y Papert, que dirigían un laboratorio del MIT dedicado a la estrategia de manipulación de símbolos y por lo tanto en competencia con los proyectos del perceptrón, comenzaron a hacer circular borradores de un libro que atacaba la idea del perceptrón. En el libro ellos hacían clara su postura científica:

Los perceptrones se han publicitado ampliamente como máquinas de «reconocimiento de patrones» o de «aprendizaje», y como tales se han discutido en gran número de libros, artículos de revista y «reportes» voluminosos. La mayor parte de estos escritos carece de valor científico.[21]

Pero su ataque fue también una cruzada filosófica. Ellos afirmaban con razón que la tradicional confianza en la reducción a primitivos lógicos había sido desafiada por un nuevo holismo:

Estos dos autores (primero independientemente y luego juntos) quedaron envueltos en algo así como una compulsión terapéutica: disipar lo que temíamos que fueran las primeras sombras de un error «holístico» o «gestáltico» que nos amenazaba con enredar los campos de la ingeniería y la inteligencia artificial tal como antes había enredado el de la biología y la psicología.[22]

Estaban en lo cierto. Las redes neuronales pueden permitir, aunque no necesariamente, una interpretación de sus nodos ocultos* en términos de rasgos que un ser humano reconocería y usaría para resolver el problema. Pese a que la modelización con redes neuronales no está comprometida con ninguna perspectiva, se puede demostrar que la asociación no *requiere* que los nodos ocultos sean interpretables. Los holistas como Rosenblatt alegremente suponían que los nodos individuales o los patrones de nodos no señalaban rasgos fijos del dominio.

Minsky y Papert estaban así tan empeñados en eliminar toda la competencia y tan seguros de la tradición atomística que va de Descartes al primer Wittgenstein, que su libro sugiere mucho más de lo que demuestra. Ellos se abocan a analizar la capacidad de un perceptrón de una sola capa,** ignorando por completo en la parte matemática de su libro los capítulos de Rosenblatt sobre las máquinas de múltiples niveles y su prueba de la convergencia de un algoritmo de aprendizaje probabilístico basada en la propagación hacia atrás*** de los errores.[23] De acuerdo con Rumelhart y McClelland,

Minsky y Papert se abocaron a mostrar qué funciones pueden o no pueden ser computadas por máquinas [de una sola capa]. Demostraron, en particular, que esos perceptrones son incapaces de calcular funciones

* Los nodos ocultos son nodos que no detectan en forma directa el insumo a la red ni constituyen su salida. Están, sin embargo, directa o indirectamente vinculados por conexiones de fuerza ajustable a los nodos que detectan los insumos y a los que constituyen su salida.

** Un perceptrón de una sola capa no tiene nodos ocultos, mientras que las redes de varios niveles sí.

*** La propagación de los errores hacia atrás requiere una computación recursiva, comenzando por los nodos de salida, a los efectos de cambiar las fuerzas de las conexiones en base a la diferencia entre la salida deseada y la salida efectivamente producida por los insumos. Durante el aprendizaje se ajustan entonces los pesos para reducir esa diferencia.

matemáticas tales como la paridad (con número par o impar de nodos en la retina) o la función topológica de conexidad (si todos los puntos positivos están conectados a otros que también lo están, ya sea directamente o por medio de otros puntos también positivos) sin hacer uso de un número absurdamente grande de predicados. El análisis es harto elegante y demuestra la importancia de una estrategia matemática para analizar los sistemas computacionales.[24]

Pero las implicancias del análisis son muy limitadas. Prosiguen Rumelhart y McClelland:

> Esencialmente aunque Minsky y Papert estuvieron finamente acertados en sus análisis del *perceptrón de una sola capa*, los teoremas no se aplican a sistemas que sean apenas un poco más complejos. En particular, no se aplican a los sistemas de capas múltiples ni a sistemas que admiten bucles de retroalimentación.[25]

Sin embargo, en la conclusión de *Perceptrons*, cuando Minsky y Papert se hacen la pregunta: ¿Han considerado los perceptrones con muchas capas?, ellos dan la impresión, mientras retóricamente dejan la pregunta abierta, de haberlos considerado:

> Bien, hemos considerado las máquinas de Gamba, que pueden ser descritas como «dos capas de perceptrón». No hemos encontrado (por haberlo pensado o por haber estudiado la literatura) ninguna otra clase interesante de máquina de capas múltiples, al menos ninguna cuyos principios estén en apariencia relacionados significativamente con los del perceptrón. [C]onsideramos que es un problema de investigación importante elucidar (o rechazar) nuestro juicio intuitivo de que la extensión es estéril.[26]

Su ataque contra el pensamiento gestáltico en IA triunfó más allá de sus sueños más descabellados. Sólo unos pocos no apreciados, entre ellos, Stephen Grossberg, James A. Anderson y Teuvo Kohonen, asumieron el «importante problema de investigación». Ciertamente, casi todos en IA supusieron que las redes neuronales se habían mandado a descansar para siempre. Rumelhart y McClelland señalan:

> El análisis de Minsky y Papert sobre las limitaciones de los perceptrones de un nivel, añadido a algunos de los éxitos iniciales de la estrategia del

procesamiento simbólico en inteligencia artificial, fue suficiente para sugerir a gran número de estudiosos en el campo que no había futuro en dispositivos análogos al perceptrón para la inteligencia artificial y la psicología cognitiva.[27]

Pero ¿era esto suficiente? Ambas estrategias habían producido algunos trabajos promisorios y algunas promesas infundadas.[28] Era aún prematuro cerrar las cuentas en cualquiera de las dos estrategias. Pero algo había en el libro de Minsky y Papert que golpeó una cuerda resonante. Parecía que los estudiosos de la IA compartían el prejuicio casi religioso contra el holismo que motivaba ese ataque. Se puede observar la fuerza de la tradición, por ejemplo, en el artículo de Newell y Simon sobre los sistemas físicos de símbolos. El artículo comienza con la hipótesis científica de que la mente y la computadora son inteligentes por virtud de la manipulación de símbolos discretos, pero finaliza con una revelación: «El estudio de la lógica y las computadoras nos ha revelado que la inteligencia reside en sistemas físicos de símbolos».[29]

El holismo no podía competir con convicciones filosóficas tan intensas. Rosenblatt quedó desacreditado junto con centenares de grupos de investigación en redes menos responsables que su obra había impulsado. Sus fondos de financiación se secaron y tuvo problemas para publicar sus trabajos. Para 1970, en lo que concierne a la IA, las redes neuronales estaban muertas. En su historia de la IA, Newell dice que la cuestión de los símbolos versus los números «ciertamente no está viva ahora y no lo ha estado por mucho tiempo.»[30] Rosenblatt no es siquiera mencionado en las historias de la IA de John Haugeland o de Margaret Boden.[31]

Pero achacar la derrota de los conexionistas a un prejuicio antiholístico es demasiado simple. Había formas más profundas en que los supuestos filosóficos influían sobre la intuición y llevaban a una sobreestimación de la importancia de los resultados del procesamiento simbólico. En aquel entonces daba la impresión de que la gente del perceptrón tenía que habérselas con una cantidad inmensa de análisis matemático y de cálculo para resolver aunque más no fuera el problema más simple de reconocimiento de patrones, tal como discriminar entre líneas horizontales y verticales en el campo perceptual, mientras que la estrategia de la manipulación simbólica había resuelto sin dificultades problemas duros del conocimiento, tales como la prueba de

teoremas lógicos y la resolución de enigmas combinatorios. Aun más importante, parecía que en función del poder computacional de que se disponía en la época, los investigadores de redes neuronales sólo podrían hacer neurociencia y psicología especulativa, mientras que los programas simples de los representacionistas simbólicos encontraban la forma de ser útiles. Detrás de esta forma de ponderar la situación yacía el supuesto de que el pensamiento y el reconocimiento de patrones eran dos dominios distintos y que el pensamiento era el más importante de los dos. Como veremos más adelante en nuestra discusión del problema del conocimiento de sentido común, ver las cosas de esta forma es ignorar el papel preeminente de la discriminación de patrones en la experiencia humana y también el trasfondo de comprensión de sentido común que está presupuesto en el pensamiento cotidiano del mundo real. Tener en cuenta este trasfondo bien puede requerir del reconocimiento de patrones.

Esta reflexión nos lleva de nuevo a la tradición filosófica. No fueron sólo Descartes y sus descendientes quienes se situaron del lado del procesamiento de la información, sino toda la tradición occidental. De acuerdo con Heidegger, la filosofía tradicional se define desde el comienzo por su interés en los hechos del mundo, mientras «pasa por encima» el mundo como tal.[32] Esto significa que la filosofía ha ignorado o distorsionado sistemáticamente desde el comienzo el contexto cotidiano de la actividad humana.[33]

La rama de la tradición filosófica que desciende de Sócrates a través de Platón, Descartes, Leibniz y Kant hasta la IA convencional da por sentado, por añadidura, que comprender un dominio consiste en poseer una teoría de ese dominio. Una teoría formula las relaciones entre los elementos objetivos, *independientes del contexto* (simples, primitivos, rasgos, atributos, factores, puntos de datos, indicios, etc) en términos de principios abstractos (leyes de cobertura, reglas, programas, etc).

Platón sostenía que, en dominios teoréticos tales como las matemáticas y quizás la ética, los pensadores aplican reglas o teorías explícitas, independientes del contexto, que han aprendido en otra vida, fuera del mundo cotidiano. Una vez aprendidas, esas teorías operan en este mundo controlando la mente del pensador, sea él consciente de ello o no. La visión de Platón no se aplica a las habilidades cotidianas sino a los dominios en los que hay conocimiento a priori. El éxito de la teoría en las ciencias

naturales, sin embargo, reforzó la idea de que en un dominio ordenado debe haber algún conjunto de elementos independientes del contexto y algunas relaciones abstractas entre esos elementos que den cuenta del orden de ese dominio y de la habilidad del hombre para actuar inteligentemente en él. Leibniz generalizaba crudamente de esta forma la concepción racionalista de todas las formas de la actividad inteligente, incluso la práctica cotidiana:

> [L]as observaciones más importantes y los talentos de habilidad en toda clase de prácticas y profesiones no se han escrito todavía. Esto lo prueba la experiencia cuando pasando de la teoría a la práctica deseamos llevar algo a cabo. *Por supuesto, también podemos describir esta práctica, dado que ella es en el fondo sólo otra teoría más compleja y particular...* [énfasis añadido][34]

La estrategia del procesamiento simbólico de la información gana su seguridad del hecho de transferir a todos los dominios los métodos que han desarrollado los filósofos y que son exitosos en las ciencias naturales. Dado que, en esta perspectiva, todos los dominios han de ser formalizables, la forma de hacer IA en cualquier área es obviamente encontrar los elementos y principios independientes del contexto y basar una representación simbólica formal en este análisis teorético. En esta vena Terry Winograd describe su trabajo en IA en términos tomados de las ciencias físicas:

> Estamos preocupados por desarrollar un formalismo, o «representación» con el cual describir el conocimiento. Buscamos los «átomos» y las «partículas» de que está hecho y las «fuerzas» que actúan en él. [35]

Sin duda, las teorías sobre el universo a menudo se construyen gradualmente modelando sistemas relativamente simples y aislados y luego tornando el modelo gradualmente más complejo e integrándolo a otros modelos de otros dominios. Esto es posible porque todos los fenómenos son presumiblemente el resultado de las relaciones legaliformes entre los que Minsky y Papert llaman «primitivos estructurales». Dado que en IA nadie *argumenta* en favor de la reducción atomística, parece que los estudiosos de IA simplemente presuponen que la abstracción de elementos fuera

de su contexto cotidiano, que define a la filosofía y funciona en las ciencias naturales, también debe funcionar en IA. Este supuesto bien puede dar cuenta de la forma en que la hipótesis del sistema físico de símbolos se tornó tan rápidamente en una revelación y de la facilidad con la que el libro de Minsky y Papert triunfó contra el holismo del perceptrón.

Mientras enseñaba filosofía en el MIT a mediados de la década de 1960, uno de nosotros —Hubert— pronto debió sumergirse en el debate sobre las posibilidades de la IA. Era obvio que investigadores como Newell, Simon y Minsky eran los herederos de la tradición filosófica. Pero dadas las conclusiones del último Wittgenstein y del primer Heidegger, eso no parecía ser de buen augurio para el programa de investigación reduccionista. Estos dos pensadores habían cuestionado la tradición misma en que se basaba el procesamiento de información simbólica. Ambos eran holistas, ambos estaban impactados por la importancia de las prácticas cotidianas y ambos sostenían que no se podía tener una teoría del mundo de la cotidianidad.

Es una de las ironías de la historia intelectual que el devastador ataque de Wittgenstein a su propio *Tractatus*, sus *Investigaciones filosóficas*,[36] se publicara en 1953, justo cuando la IA se situaba en la tradición abstracta y atomista que él había atacado. Tras escribir su *Tractatus*, Wittgenstein pasó años haciendo lo que él llamaba fenomenología[37], buscando en vano los hechos atómicos y los objetos básicos que su teoría requería. Terminó por abandonar su *Tractatus* y toda la filosofía racionalista. Argumentaba que el análisis de las situaciones cotidianas en términos de hechos y reglas (que es donde la mayoría de los filósofos tradicionales y los investigadores de IA piensan que debería comenzar la teoría) es sólo significativo en algunos contextos y para algunos propósitos. De tal modo, los elementos escogidos ya reflejan los objetivos y propósitos para los que fueron puestos en mira. Cuando procuramos encontrar elementos independientes del contexto y libres de propósitos, como debemos hacerlo si pretendemos encontrar los símbolos primitivos con que alimentar una computadora, estamos tratando en realidad de liberar aspectos de nuestra experiencia de la misma organización pragmática que hace posible usarlos inteligentemente al enfrentar los problemas de la vida diaria.

En las *Investigaciones filosóficas* Wittgenstein criticó directamente el atomismo lógico de su *Tractatus*:

«¿Qué hay detrás de la idea de que los nombres significan en realidad cosas primarias?», decía Sócrates en el *Teéteto*: «Si no me equivoco, he escuchado que algunos decían esto: no hay definición de los elementos primarios —por así decirlo— a partir de los cuales nosotros y todas las cosas estamos compuestos. Pero así como lo que está compuesto por esos elementos es en sí complejo, del mismo modo los nombres de los elementos devienen lenguaje descriptivo al reunirse». Tanto los 'individuos' de Russell como mis 'objetos' (*Tractatus Logico-Philosophicus*) eran de esos elementos primarios. Pero ¿cuáles son las partes constituyentes simples de que está compuesta la realidad? No tiene en absoluto sentido hablar de las 'partes simples de una silla.'[38]

Ya en la década de 1920 Martin Heidegger había reaccionado del mismo modo contra su mentor, Edmund Husserl, quien se consideraba la culminación de la tradición cartesiana, siendo por ello el abuelo de la IA.[39] Husserl argumentaba que un acto de conciencia o noesis no capta en sí a un objeto; más bien, el acto posee intencionalidad (direccionalidad) sólo en virtud de una «forma abstracta» o significado en el noema correlacionado con ese acto.[40]

Este significado o representación simbólica, tal como Husserl lo concebía, es una entidad compleja que tiene una dura tarea que realizar. En *Ideas Pertaining to a Pure Phenomenology*,[41] Husserl trató de explicar arduamente cómo el noema hace su trabajo. Los «sentidos de predicado» proporcionan la referencia, y al igual que el *Sinne* fregeano, tienen la notable propiedad de captar las propiedades atómicas de los objetos. Esos predicados se combinan en «descripciones» complejas de objetos complejos, como en la teoría de las descripciones de Russell. Para Husserl, que estaba cerca de Kant en este punto, el noema consiste en una jerarquía de reglas estrictas. Dado que Husserl concebía la inteligencia como una actividad determinada por el contexto y orientada hacia objetivos, la representación mental de cualquier tipo de objetos tenía que proporcionar un contexto o un «horizonte» de expectativas o «predelineamientos» para estructurar los datos entrantes: «una regla que gobierna otras posibles conciencias del [objeto]; posibles, en el sentido de ejemplificar tipos esencialmente predelineados.»[42] El noema debe contener una regla que describe todos los rasgos que pueden esperarse con certidumbre al explorar cierto tipo de objeto, rasgos que permanecen «inviolablemente

los mismos: en tanto como la objetividad siga siendo intencionada como esta y de esta clase.»[43] La regla debe incluir también predelineamientos de las propiedades que son rasgos posibles, pero no necesarios, de este tipo de objeto: «En vez de un sentido completamente determinado, siempre hay, por lo tanto, un *marco de sentido vacío...*»[44]

En 1973 Marvin Minsky propuso una nueva estructura de datos, notoriamente parecida a la de Husserl, para representar el conocimiento cotidiano:

> Un marco* es una estructura de datos que representa una situación estereotipada, como la de estar en cierta clase de habitación, o la de ir a la fiesta de cumpleaños de un niño.
>
> Podemos concebir un marco como una red de nodos y relaciones. Los niveles superiores de un marco son fijos y representan cosas que siempre son verdad respecto de la situación supuesta. Los niveles inferiores tienen muchos terminales, ranuras que se deben llenar de instancias específicas o datos. Cada terminal puede especificar las condiciones que deben satisfacer sus asignaciones.
>
> Gran parte del poder fenomenológico de la teoría reside en la inclusión de expectativas y otras clases de supuestos. *Los terminales de un marco están ya llenos con asignaciones «por defecto»*.[45]

En el modelo del marco de Minsky, el «nivel superior» es una versión desarrollada de lo que en la terminología de Husserl permanece «inviolablemente lo mismo» en la representación, y los predelineamientos de Husserl devienen «asignaciones por defecto», rasgos adicionales que se pueden esperar en condiciones normales. El resultado es un paso de avance en las técnicas de IA más allá de un modelo pasivo de procesamiento de información, hacia un modelo que intenta tomar en cuenta las interacciones entre quien conoce y el mundo. La tarea de la IA converge entonces con la de la fenomenología trascendental. Ambas deben encontrar, en situaciones cotidianas, los marcos construidos a partir de un conjunto de predicados primitivos y sus relaciones formales.

Antes que Wittgenstein, Heidegger desarrolló, en respuesta a Husserl, una descripción fenomenológica del mundo cotidiano y

* *Frame* en el original. En la comunidad computacional de habla española a menudo se utiliza el vocablo inglés; lo traducimos sin embargo como «marco» porque ése es el concepto de uso en la tradición filosófica [T.].

de objetos cotidianos como sillas y martillos. Al igual que Wittgenstein, encontró que el mundo cotidiano no se podía representar mediante un conjunto de elementos independientes del contexto. Fue Heidegger quien impulsó a Husserl a encarar precisamente este problema, al señalar que hay otras formas de «encontrarse» con las cosas, aparte de relacionarlas con objetos definidos mediante un conjunto de predicados. Cuando usamos una pieza de equipamiento como un martillo, decía Heidegger, actualizamos una habilidad (que no necesita estar representada en la mente) en el contexto de un nexo socialmente organizado de equipamiento, propósitos y papeles humanos (que no necesitan estar organizados como un conjunto de hechos). Este contexto, o mundo, y nuestras formas cotidianas de lidiar hábilmente con él, lo que Heidegger llama «circunspección», no es algo que *pensamos* sino parte de nuestra socialización que constituye la forma en que *somos*. Concluía Heidegger:

> El contexto se puede considerar formalmente en el sentido de un sistema de relaciones. Pero [e]l contexto fenoménico de esas «relaciones» y «relata» es tal que resiste cualquier clase de funcionalización matemática; tampoco es algo que sea meramente pensado, estipulado por primera vez en un «acto de pensamiento». Son más bien relaciones en las que la circunspección interesada ya habita de antemano.[46]

Esto define la divergencia en los caminos de Husserl y la IA por un lado y de Heidegger y el último Wittgenstein por el otro. La pregunta crucial es entonces: ¿Puede haber una teoría del mundo cotidiano como la que han sostenido siempre los filósofos racionalistas? ¿O el conocimiento contextual de sentido común es más bien una combinación de habilidades, prácticas, discriminaciones, etc., que no son estados intencionales y, por lo tanto, *a fortiori*, no poseen ningún contenido representacional explicable en términos de elementos y reglas?

Haciendo una movida que pronto se hizo familiar en los círculos de la IA, Husserl procuró soslayar el problema que Heidegger había planteado. Husserl alegaba que el mundo, el trasfondo de sentido, el contexto cotidiano, era meramente un sistema muy complejo de hechos correlativo a un sistema complejo de creencias, a las cuales, dado que tenían condiciones de verdad, él llamó validaciones. Sostenía que en principio se podía suspender la confianza en el

mundo y alcanzar una descripción desapegada del sistema de creencias humano. Se podía entonces completar la tarea que había estado implícita en la filosofía desde Sócrates: se podían hacer implícitos los principios y creencias subyacentes a toda conducta inteligente. Como lo expresaba Husserl,

> [I]ncluso el trasfondo... del cual siempre estamos concurrentemente conscientes pero que es momentáneamente irrelevante y permanece por completo inadvertido, funciona de acuerdo con sus validaciones implícitas.[47]

Dado que creía firmemente que el trasfondo compartido podría tornarse explícito como sistema de creencias, Husserl se adelantaba a su tiempo suscitando el tema de la posibilidad de la IA. Tras discutir la posibilidad de que un sistema formal axiomático pudiera describir la experiencia y señalando que un sistema así de axiomas y primitivos (al menos como lo conocemos en geometría) no podría describir formas cotidianas tales como «en forma de peine» o «lenticular», Husserl dejaba abierta la cuestión de si esos conceptos cotidianos se podían después de todo formalizar. Esto era como dejar abierto el problema de la IA respecto de si es posible formalizar la física del sentido común. Retomando el sueño de Leibniz de la mathesis de toda la experiencia, Husserl agregaba:

> La cuestión importante es si no podría haber un procedimiento de idealización que sustituya por ideales puros y estrictos los datos intuidos y que pudiera servir como recurso básico de la mathesis de la experiencia.[48]

Pero, como Heidegger había predicho, la tarea de describir una concepción teorética completa de la vida cotidiana demostró ser mucho más ardua de lo que se esperaba inicialmente. El proyecto de Husserl se extravió en serios problemas, y hay signos de que el de Minsky también. Después de 25 años de intentar determinar los componentes de la representación subjetiva de los objetos cotidianos, Husserl encontró que debía incluir más y más elementos de la comprensión subjetiva de sentido común del mundo de la vida cotidiana:

> Con certeza, incluso las tareas que se presentan cuando tomamos tipos simples de objetos como signos restringidos demuestran ser extre-

madamente complicadas, y a menudo llevan a disciplinas extensivas cuando penetramos más profundamente. Este es el caso, por ejemplo, con los objetos espaciales como tales (para no decir nada de la naturaleza), del ser psicofísico y de la humanidad como tal, de la cultura como tal.[49]

Husserl hablaba de la «pesada concretidad»[50] del noema y de su «tremenda complejidad,»[51] y concluía con tristeza, a la edad de setenta y cinco años, que él había sido un principiante perpetuo y que la fenomenología era una «tarea infinita.»[52]

En su ensayo «A Framework for Representing Knowledge», hay indicios de que Minsky se ha embarcado en la misma «tarea infinita» que en su momento desbordó a Husserl:

Construir una base de conocimientos es un problema de investigación intelectual mayor. Aún sabemos demasiado poco sobre los contenidos y la estructura del conocimiento de sentido común. Un sistema de sentido común «mínimo» debe «saber» algo sobre causa y efecto, tiempo, propósito, localización, proceso y tipos de conocimiento. Necesitamos un esfuerzo de investigación epistemológica muy serio en esta área.[53]

Aun para un aprendiz de filosofía contemporánea, la ingenuidad y la confianza de Minsky resultan asombrosas. La fenomenología de Husserl fue justamente uno de esos esfuerzos de investigación. Ciertamente, los filósofos desde Sócrates hasta Wittgenstein, pasando por Leibniz, llevaron adelante investigación epistemológica seria en esta área durante dos mil años sin éxito apreciable.

A la luz del fracaso de Wittgenstein y de la devastadora crítica de Husserl, uno de nosotros (Hubert) predijo problemas al procesamiento simbólico de la información. Como señala Newell en su historia de la IA, esa advertencia fue ignorada:

La objeción intelectual principal de Dreyfus es que el análisis del contexto de la acción humana en elementos discretos está condenada a fracasar. Esta objeción se funda en la filosofía fenomenológica. Desafortunadamente, esto no parece ser de importancia en lo que a la IA se refiere. Las respuestas, refutaciones y análisis que han sucedido a los artículos de Dreyfus sencillamente no se han interesado en la cuestión; cuestión que sería un aspecto novedoso si se llegara a poner en primer plano.[54]

El problema, sin embargo, no tardó en aparecer en primer plano, pues el mundo cotidiano se tomó venganza tanto de la IA como de la filosofía tradicional. Tal como lo vemos, el programa de investigación lanzado por Newell y Simon ha atravesado tres etapas de diez años. De 1955 a 1965 dos temas de investigación, la representación y la búsqueda, dominaron el campo denominado entonces «simulación cognitiva». Newell y Simon demostraron, por ejemplo, cómo resuelve una computadora una clase de problemas ateniéndose al principio heurístico de búsqueda general conocido como análisis de medios y fines; esto es, utilizar cualquier operación disponible que reduzca la distancia entre la descripción de la situación actual y la descripción del objetivo. Luego abstrajeron esta técnica heurística y la incorporaron en su General Problem Solver (GPS).

La segunda etapa (1965-1975), liderada por Marvin Minsky y Seymour Papert en el MIT, tenía que ver con los hechos y reglas a representar. La idea consistía en desarrollar métodos para tratar sistemáticamente el conocimiento en dominios aislados llamados «micromundos». Algunos de los programas famosos escritos hacia 1970 en el MIT fueron SHRDLU de Terry Winograd, que podía obedecer órdenes dadas en un subconjunto de lenguaje natural a propósito de un «mundo de bloques» simplificado, el programa del problema de la analogía de Thomas Evan, el de análisis escénico de David Waltz y el programa de Patrick Winston, que podía aprender conceptos a través de ejemplos.

Se esperaba que los micromundos, restringidos y aislados, fueran cada vez más realistas y que se combinaran de manera tal de parecerse a la comprensión de sentido común. Pero los investigadores confundieron los dos dominios que, siguiendo a Heidegger, distinguiríamos como «universo» y «mundo». Un conjunto de hechos interrelacionados puede constituir un universo, como el universo físico, pero no constituye un mundo. Este último, como el mundo de los negocios, el mundo del teatro o el de los físicos, es un cuerpo organizado de objetos, propósitos, habilidades y prácticas sobre la base del cual adquieren sentido las actividades humanas. Para apreciar la diferencia, se puede contrastar el universo físico sin significado con el mundo pleno de significados de la disciplina de la física. El mundo de la física, el mundo de los negocios y el mundo del teatro tienen sentido sólo contra un trasfondo de temas humanos comunes. Son elaboraciones locales

del mundo de sentido común que todos compartimos. Es decir, los submundos no están relacionados como sistemas físicos aislables a un sistema mayor que ellos componen, sino que más bien son elaboraciones locales de un todo al que ellos presuponen. Los micromundos no son mundos sino dominios aislados sin significación, y gradualmente ha llegado a ser evidente que no hay forma en que pueda combinarse y extenderse hasta llegar al mundo de la vida cotidiana.

En su tercera etapa, más o menos desde 1975 hasta el presente, la IA ha estado luchando con lo que se ha dado en llamar el problema del conocimiento del sentido común. La representación del conocimiento siempre ha sido un problema central para el trabajo en IA, pero los dos períodos iniciales (la simulación cognitiva y los micromundos) se caracterizaron por el intento de eludir el problema del conocimiento de sentido común y por ver qué podía hacerse haciendo intervenir tan poco conocimiento como fuera posible. A mediados de la década de 1970, sin embargo, el problema tuvo que encararse. Se probaron sin éxito diversas estructuras de datos, como los marcos de Minsky y los argumentos* de Roger Schanks. El problema del conocimiento de sentido común ha impedido a la IA cumplir la predicción hecha por Simon hace veinte años de que «dentro de veinte años las máquinas serán capaces de realizar cualquier trabajo que pueda realizar un hombre.»[55]

De hecho, el problema del conocimiento de sentido común ha bloqueado todo progreso en la IA teorética durante la última década. Winograd ha sido uno de los primeros en ver las limitaciones de SHRDLU y de los intentos de los marcos y argumentos para extender la estrategia de los micromundos. Habiendo perdido la fe en la IA, él enseña ahora Heidegger en su curso de computación en Stanford y señala «la dificultad de formalizar el fondo de sentido común que determina qué argumentos, metas y estrategias son relevantes y cómo interactúan».[56]

Lo que sostiene a la IA en esta *impasse* es la convicción de que el problema del conocimiento de sentido común se puede resolver, ya que los seres humanos obviamente lo han resuelto. Pero bien puede que los seres humanos no usen conocimiento de sentido común en absoluto. Como lo han señalado Heidegger y Wittgenstein,

* *Scripts* [T.].

la comprensión de sentido común bien pudiera ser un saber-cómo de sentido común. Por «saber-cómo» no queremos decir reglas de procedimiento, sino saber qué hacer en un vasto número de casos especiales.[57] Por ejemplo, la física de sentido común se ha tornado algo extremadamente difícil de enunciar en un conjunto de hechos y reglas. Cuando uno lo intenta, o bien se requiere más sentido común para comprender los hechos y las reglas que se encuentran, o bien se producen fórmulas de tal complejidad que parece poco probable que se encuentren en la mente de un niño.

Hacer física teórica también requiere habilidades de fondo que pueden no ser formalizables, pero el dominio mismo se puede describir mediante leyes abstractas que no hacen referencia a esas habilidades básicas. Los investigadores de IA concluyeron erróneamente que la física de sentido común también podía expresarse como un conjunto de principios abstractos. Pero bien puede ser que el problema de encontrar una teoría de la física de sentido común sea insoluble, porque el dominio no posee estructura teorética. Jugando con líquidos y sólidos de todas clases durante años, todos los días, el niño bien podría aprender a discriminar casos prototípicos de sólidos, líquidos y cosas así y aprender respuestas hábiles típicas a comportamientos típicos en circunstancias típicas. Lo mismo puede decirse para el caso del mundo social. Si el conocimiento de fondo es una habilidad y si las habilidades se basan en patrones holísticos y no en reglas, podríamos esperar que las representaciones simbólicas sean incapaces de capturar nuestra comprensión de sentido común.

A la luz de esta *impasse*, la IA clásica, basada en símbolos, parece ser más bien un ejemplo perfecto de lo que Imre Lakatos ha llamado un programa de investigación degenerativo.[58] Como vimos, la IA comenzó auspiciosamente con el trabajo de Newell y Simon en la Rand y para fines de la década de 1960 se había convertido en un programa de investigación floreciente. Minsky predijo que «dentro de una generación el problema de crear 'inteligencia artificial' habrá sido sustancialmente resuelto».[59] Entonces, casi de repente, el campo cayó en dificultades inesperadas. Formular una teoría del sentido común demostró ser más arduo de lo que se había pensado. No sólo se trataba, como esperaba Minsky, de una cuestión de catalogar unos pocos centenares de miles de hechos. El conocimiento de sentido común se convirtió en el centro de las preocupaciones. El humor de

Minsky cambió por completo en cinco años. Dijo a un reportero que «el problema de la IA es uno de los más difíciles que jamás haya abordado la ciencia».[60]

La tradición racionalista debió colocarse finalmente frente a una prueba empírica y fracasó. La idea de producir una teoría formal, atomística, del mundo cotidiano del sentido común cayó exactamente en las dificultades que Heidegger y Wittgenstein habían descubierto. La intuición de Frank Rosenblatt de que sería desesperanzadoramente difícil formalizar el mundo y dar así una especificación formal de la conducta inteligente ha sido vindicada. Su reprimido programa de investigación (utilizar la computadora para instanciar un modelo holístico de un cerebro idealizado), que nunca había sido realmente refutado, volvió a ser una opción viva.

En las reseñas periodísticas de la historia de la IA, detractores anónimos denigran a Rosenblatt como si fuera un vendedor de aceite de serpiente:

> Los investigadores actuales recuerdan que Rosenblatt se había dado a insistentes y extravagantes afirmaciones sobre la performance de esta máquina. «El fue el sueño de un agente de prensa,» dice un científico, «un verdadero hechicero. Si lo escuchábamos a él, el perceptrón era capaz de cosas fantásticas. Y puede que lo fuera. Pero no se lo puede probar por el trabajo que Frank hizo.»[61]

De hecho, él tenía más claras las capacidades y limitaciones de los diversos tipos de perceptrón de lo que Simon y Minsky tenían claro las de los programas simbólicos.[62] Ahora está siendo rehabilitado. David Rumelhart, Geoffrey Hinton y James McClelland reflejan esta nueva apreciación de su obra pionera:

> El trabajo de Rosenblatt era muy polémico para su época, y los modelos específicos que propuso no satisfacían lo que él esperaba de ellos. Pero esta concepción del sistema humano de procesamiento de la información como sistema dinámico, interactivo y autoorganizado está en el núcleo mismo de la estrategia del PDP.[63]
>
> Los estudios de los perceptrones anticipan claramente muchos de los resultados que hoy están en uso. Se interpretó equivocadamente la crítica de Minsky y Papert como algo que destruía su credibilidad, aunque el trabajo sólo mostraba las limitaciones de la clase más simple de mecanis-

mos afines al perceptrón, sin decir nada acerca de los modelos más poderosos de múltiples capas.[64]

Los frustrados investigadores de la IA, hartos de estar pegados a un programa de investigación que Jerry Letvin caracterizó a principios de la década de 1980 como «la única paja que flota», se cambiaron al nuevo paradigma. El libro de Rumelhart y McClelland *Parallel Distributed Processing* vendió seis mil copias el día que salió al mercado, y treinta mil copias más están en prensa. Como lo dice Paul Smolensky,

> En la última media década la estrategia conexionista de modelización cognitiva ha dejado de ser un culto oscuro con unos pocos creyentes auténticos hasta convertirse en un movimiento tan vigoroso que los encuentros recientes de la Cognitive Science Society han comenzado a parecerse a rallies de fuerza conexionistas.[65]

Si las redes de múltiples capas logran cumplir su promesa, los investigadores tendrán que renunciar a la convicción de Descartes, Husserl y el primer Wittgenstein de que la única forma de producir conducta inteligente es reflejando el mundo mediante una teoría formal de la mente. Peor aún, se tendrá que renunciar a la intuición más básica que está en la fuente de la filosofía de que debe haber una teoría para cada aspecto de la realidad; esto es, que deben existir elementos y principios en términos de los cuales se puede dar cuenta de la inteligibilidad de cualquier dominio. Las redes neuronales pueden demostrar que Heidegger, el último Wittgenstein y Rosenblatt tenían razón al pensar que nos comportamos inteligentemente en el mundo sin tener una teoría de ese mundo. Si no es *necesaria* una teoría para explicar la conducta inteligente, tenemos que estar preparados para plantear la cuestión de si es a fin de cuentas *posible* tal explicación teorética en los dominios de la vida cotidiana.

Una vez que sus redes han sido entrenadas para ejecutar una tarea, los modeladores de redes neuronales, influenciados por una IA manipuladora de símbolos, están invirtiendo un esfuerzo considerable para hallar los rasgos representados por cada nodo individual o conjunto de nodos. Los resultados hasta ahora son equívocos. Consideremos las redes de Hinton para aprender conceptos por medio de representaciones distribuidas.[66] La red se

puede entrenar para codificar relaciones en un dominio que los seres humanos conceptualizan en términos de rasgos, sin dar a la red los rasgos que los humanos utilizan. Hinton produce ejemplos de casos en los cuales algunos nodos de la red entrenada se pueden interpretar en correspondencia con los rasgos que los seres humanos perciben, aunque los nodos sólo a grandes trazos corresponden a esos rasgos. La mayor parte de los nodos, sin embargo, no se puede interpretar en absoluto semánticamente. Un rasgo utilizado en la representación simbólica está presente o no lo está. En la red, sin embargo, aunque ciertos nodos se encuentran más activos cuando está presente cierto rasgo en un dominio, la suma de actividad no sólo varía con la presencia o ausencia de ese rasgo, sino que se encuentra afectada por la presencia o ausencia de otros rasgos.

Hinton ha abordado un dominio (las relaciones familiares) que los seres humanos construyen precisamente en términos de los rasgos que los humanos normalmente perciben, tales como generación y nacionalidad. Hinton analiza entonces esos casos en los que, comenzando con ciertas conexiones de fuerzas inicialmente al azar, algunos nodos pueden, después del aprendizaje, interpretarse como representativos de esos rasgos. Los cálculos que utilizan el modelo de Hinton, sin embargo, muestran que incluso su red parece aprender sus asociaciones para algunas fuerzas de conexión iniciales al azar sin ningún uso obvio de esos rasgos de la vida cotidiana.

En un sentido muy limitado, cualquier red de capas múltiples bien entrenada se puede interpretar en términos de rasgos, no de los rasgos cotidianos sino de lo que podríamos llamar rasgos altamente abstractos. Consideremos el caso simple de los niveles de unidades binarias activadas por conexiones de prealimentación*, pero no conexiones laterales o de retroalimentación. Para construir una concepción tal de una red que ha aprendido ciertas asociaciones, cada nodo que está un nivel por encima de los nodos de entrada se puede interpretar, sobre la base de sus conexiones con él, como si detectara cuándo está presente un patrón entre un conjunto de patrones de entrada. (Algunos patrones serán los que se usan en el entrenamiento, otros no habrán sido utilizados antes.) Si se da un nombre inventado al conjunto de patrones de

* *Feed-forward* [T.].

entrada que detecta un nodo en particular (que ciertamente no tendría un nombre en nuestro vocabulario), se puede interpretar que el nodo detecta el rasgo altamente abstracto así denominado. De allí, cada nodo que esté un nivel por encima del nivel de entrada puede caracterizarse como un detector de rasgos. En forma parecida, cada nodo que esté un nivel por encima de esos nodos puede interpretarse como si detectara un rasgo de un orden más alto, definido como la presencia de un conjunto especificado de patrones entre el primer nivel de detectores de rasgos. Y así hacia arriba en la jerarquía.

El hecho de que se pueda dar cuenta de la inteligencia, definida como el conocimiento de cierto conjunto de asociaciones apropiadas a un dominio, en términos de relaciones entre rasgos altamente abstractos de un dominio de habilidades, sin embargo, no preserva la intuición racionalista de que esos rasgos explicativos deben capturar la estructura esencial del dominio de modo que se pueda basar sobre ellos una teoría. Si se enseña a la red una asociación más de un par de entrada-salida (donde el insumo anterior al entrenamiento produjo una salida diferente de la que se debe aprender), se deberá cambiar al menos la interpretación de algunos de los nodos. De este modo, los rasgos que algunos de los nodos entresacaron antes de la última instancia de entrenamiento resultarán no haber sido rasgos estructurales invariantes del dominio.

Una vez que se abandona la estrategia filosófica de la IA clásica y se acepta el reclamo ateorético de la modelización de redes neuronales, subsiste una cuestión: ¿Cuánta inteligencia cotidiana puede esperarse que capture una red semejante? Los investigadores de la IA clásica se apresuran a señalar —como ya lo había hecho Rosenblatt— que los modelizadores de redes neuronales hasta ahora difícilmente hayan tratado resoluciones de problemas de dificultad creciente. Esta respuesta, sin embargo, nos recuerda demasiado la forma en que los manipuladores simbólicos de los sesenta respondían a la crítica de que sus programas eran pobres para la percepción de patrones. La vieja lucha continúa entre los intelectualistas, que piensan que porque pueden hacer lógica independiente del contexto ellos poseen la clave del conocimiento cotidiano, pero son flojos comprendiendo la percepción, y los gestálticos, que tienen los rudimentos de un concepto de percepción pero no del conocimiento cotidiano.[67] Se

podría pensar, usando la metáfora del cerebro derecho e izquierdo, que quizás el cerebro o la mente usa cada estrategia cuando resulta adecuado. El problema sería entonces cómo combinar las estrategias. No se puede conmutar para uno u otro lado, pues como lo advirtieron Heidegger y los gestálticos, el fondo pragmático juega un papel crucial para determinar la relevancia (incluso en la lógica cotidiana y en la resolución de problemas), y los expertos en cualquier campo, incluso la lógica, captan las operaciones en términos de sus parecidos funcionales.

Todavía es prematuro pensar en combinar las dos estrategias, dado que por el momento ninguna ha logrado lo suficiente como para asentarse en terreno seguro. La estrategia del sistema físico de símbolos parece estar fallando sencillamente porque es falso suponer que debe haber una teoría de cada dominio. La modelización con redes neuronales, sin embargo, no está comprometida con esta ni con ninguna otra presunción filosófica.

No obstante, construir una red interactiva lo suficientemente parecida a la que nuestro cerebro ha desarrollado puede ser demasiado arduo. El problema del conocimiento de sentido común, que ha bloqueado el progreso de las técnicas de representación simbólica por quince años, puede estar asomando en el horizonte de las redes neuronales, aunque los investigadores todavía no lo reconozcan. Todos los modelizadores de redes neuronales están de acuerdo en que para que una red sea inteligente debe ser capaz de generalizar; esto es, dados suficientes ejemplos de insumos asociados con una salida particular, debe asociar luego insumos del mismo tipo con salidas de la misma clase. Surge la cuestión, sin embargo: ¿Qué es lo que se cuenta como del mismo tipo? El diseñador de la red tiene en mente una definición específica del tipo requerido para una generalización razonable y cuenta como si fuera un éxito cuando la red generaliza otras instancias del mismo tipo. Pero cuando la red produce una asociación inesperada, ¿puede decirse que se ha equivocado en la generalización? Lo mismo podría decirse que la red ha venido trabajando en torno a una definición diferente del tipo en cuestión y que esa diferencia acaba de revelarse. (Todas las cuestiones del tipo «continúa esta secuencia» que se encuentran en los tests de inteligencia poseen realmente más de una respuesta posible, pero la mayoría de los seres humanos comparte un sentido de lo que es simple y razonable, y por lo tanto aceptable.)

Los modelizadores de redes neuronales intentan evitar esta ambigüedad y hacer que la red produzca generalizaciones «razonables» considerando sólo una familia de generalizaciones posibles preespecificada, esto es, transformaciones permitidas que contarán como generalizaciones aceptables (el espacio de hipótesis). Estos modelizadores pretenden entonces diseñar la arquitectura de sus redes de manera tal que ellas transformen los insumos en salidas sólo en formas que se encuentran en el espacio de hipótesis. La generalización sólo será posible entonces en los términos del diseñador. Mientras que unos pocos ejemplos serán insuficientes para identificar unívocamente los miembros apropiados del espacio de hipótesis, después de suficientes ejemplos solo una hipótesis dará cuenta de todos los ejemplos. La red habrá aprendido entonces el principio de generalización apropiado. Es decir, todo insumo ulterior producirá lo que, desde el punto de vista del diseñador, es la salida apropiada.

El problema aquí es que el diseñador ha determinado, por medio de la arquitectura de la red, que ciertas generalizaciones posibles nunca se encuentren. Todo esto es bueno y está bien para problemas de juguete en los que no hay problemas respecto de lo que constituye o no una buena generalización; pero en situaciones de la vida real buena parte de la inteligencia humana consiste en generalizar modos que son apropiados a un contexto. Si el diseñador restringe la red a una clase predefinida de respuestas apro-piadas, la red estará exhibiendo la inteligencia construida dentro de ella por el diseñador para ese contexto, pero no tendrá el sentido común que la habilitará para adaptarse a otros contextos, como lo haría una inteligencia genuinamente humana.

Quizás una red deba compartir su tamaño, su arquitectura y su configuración de conexiones iniciales con el cerebro humano, si es que ha de compartir nuestro sentido de generalización apropiada. Ha de aprender de sus propias «experiencias» para hacer asociaciones que se asemejen a las humanas, y no tanto ser enseñada a hacer asociaciones que son específicas de su entrenador; una red debe compartir también nuestro sentido de propiedad de un comportamiento, y esto significa que debe compartir nuestras necesidades, deseos y emociones y poseer un cuerpo parecido al humano con movimientos físicos, habilidades y vulnerabilidad a la violencia igualmente apropiados.

Si Heidegger y Wittgenstein están en lo cierto, los seres humanos son mucho más holísticos que las redes neuronales. La inteligencia tiene que estar motivada por propósitos en el organismo y por objetivos seleccionados por el organismo a partir de una cultura en marcha. Si la unidad mínima de análisis es la de todo un organismo adaptado a todo un mundo cultural, las redes neuronales, tanto como las computadoras simbólicamente programadas, todavía tienen un largo camino por delante.

Notas

[1] Ludwig Wittgenstein, *Last Writings on the Philosophy of Psychology* (Chicago: Chicago University Press, 1982), vol. 1, 504 (66e). Traducción modificada.

[2] David E. Rumelhart y Donald A. Norman, «A Comparison of Models», *Parallel Models of Associative Memory*, ed. de Geoffrey Hinton y James Anderson (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1981), 3.

[3] Allen Newell, «Intellectual Issues in the History of Artificial Intelligence», en *The Study of Information: Interdisciplinary Messages*, editado por F. Machlup y U. Mansfield (Nueva York: Wiley, 1983), 196.

[4] Allen Newell y Herbert Simon, «Computer Science as Empirical Inquiry: Symbols and Search», reimpreso en *Mind Design*, editado por John Haugeland (Cambridge: MIT Press, 1981), 41.

[5] Ibid., 42.

[6] Thomas Hobbes, *Leviathan* (Nueva York: Library of Liberal Arts, 1958), 45.

[7] Leibniz, *Selections*, editado por Philip Wiener (Nueva York, Scribner, 1951), 18.

[8] Ibid., 20.

[9] Ibid., 10.

[10] Ludwig Wittgenstein, *Tractatus Logico-Philosophicus* (Londres, Routledge y Kegan Paul, 1960) [Traducción española: *Tractatus Logico-Philosophicus*, Madrid, Alianza, 1973].

[11] D.O. Hebb, *The Organization of Behavior* (Nueva York, Wiley, 1949).

[12] Frank Rosenblatt, «Strategic Approaches to the Study of Brain Models». *Principles of Self-Organization*, editado por H. von Foerster (Elmsford, N.Y.: Pergamon Press, 1962), 386

[13] Ibid., 387.

[14] Herbert Simon y Allen Newell, «Heuristic Problem Solving: The Next Advance in Operations Research», *Operations Research* 6 (enero-febrero de 1958):6.

[15] Ibid. Las reglas heurísticas son reglas que cuando las usan los seres humanos se dice que están basadas en la experiencia o en el juicio. Tales reglas conducen con frecuencia a soluciones plausibles de los problemas o incrementan la eficiencia de un procedimiento de resolución de problemas. Mientras que los algoritmos garantizan una solución correcta (si es que hay una) en un tiempo finito, las heurísticas sólo incrementan la probabilidad de encontrar una solución plausible.

[16] David E. Rumelhart, James L. McClelland y el PDP Research Group, en su reciente compilación de ensayos *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1 (Cambridge: MIT Press, 1986), describen el perceptrón como sigue:

"Tales máquinas consisten en lo que se llama generalmente una retina, un patrón de insumos binarios que a veces se disponen en un espacio bidimensional; un conjunto de predicados, un conjunto de unidades de umbral binario con conexiones fijas a un conjunto de unidades en la retina tal que cada predicado computa alguna función local sobre el conjunto de unidades al que está conectado; y una o más unidades de decisión, con conexiones modificables a los predicados". (p. 111)

Los autores contrastan la forma en que un modelo de procesamiento distribuido (PDP) como el perceptrón almacena información con la forma en que se almacena la información en la representación simbólica:

"En la mayoría de los modelos, el conocimiento se almacena como la copia estática de un patrón. La recuperación consiste en encontrar el patrón en la memoria a largo plazo y copiarla a un *buffer* en la memoria de trabajo. No hay diferencia real entre la información almacenada en la memoria a largo plazo y la representación activa en la memoria de trabajo. En los modelos PDP, sin embargo, éste no es el caso. En estos modelos, los patrones mismos no se almacenan. Más bien lo que se almacena son las *fuerzas de conexión* entre las unidades, que permiten recrear esos patrones." (p. 31)

"[E]l conocimiento sobre un patrón individual no se almacena en las conexiones de una unidad especial reservada para ese patrón, sino que se distribuye sobre las conexiones de un número grande de unidades de procesamiento." (p. 33)

Esta nueva noción de representación llevaba directamente a la idea de Rosenblatt de que tales máquinas serían capaces de adquirir sus habilidades mediante el aprendizaje en lugar de ser programadas con rasgos y reglas:

"[S]i el conocimiento yace [en] las fuerzas de las conexiones, el aprendizaje debe ser cuestión de encontrar las fuerzas de conexión correctas, tal que se produzcan los patrones correctos de activación en las circunstancias adecuadas. Esta es una propiedad extremadamente importante de esta clase de modelos, puesto que abre la posibilidad de que un mecanismo de procesamiento pueda aprender como resultado del afinamiento de sus conexiones, capturar las interdependencias entre las activaciones que se exponen en el curso del procesamiento." (p. 32)

[17] Frank Rosenblatt, *Mechanisation of Thought Processes: Proceedings of a Symposium held at the National Physical Laboratory* (Londres: Her Majesty's Stationery Office, 1948), vol. 1, 449.

[18] Marvin Minsky y Seymour Papert, *Perceptrons: An Introduction to Computational Geometry* (Cambridge: MIT Press, 1969), 19.

[19] Sir James Lighthill, «Artificial Intelligence: A General Survey», en *Artificial Intelligence: A Paper Symposium* (Londres, Science Research Council, 1973).

[20] Rumelhart y McClelland, *Parallel Distributed Processing*, 158.

[21] Minsky y Papert, *Perceptrons*, 4.

[22] Ibid., 19.

[23] Frank Rosenblatt, *Principles of Neurodynamics, Perceptrons and the Theory of Brain Mechanisms* (Washington, D.C.: Spartan Books, 1962), 292. Véase también:

"La adición de un cuarto nivel de unidades de trasmisión de señales, o el acoplamiento cruzado de las unidades-A de un perceptrón de tres capas, permite

la solución de problemas de generalización sobre grupos de transformación arbitrarios." (p. 576)

"En perceptrones acoplados hacia atrás, puede manifestarse una atención selectiva hacia objetos familiares en un campo complejo. También le es posible al perceptrón fijarse selectivamente en objetos que se mueven diferencialmente con respecto a su entorno." (p. 576).

[24] Rumelhart y McClelland, *Parallel Distributed Processing*, 111.

[25] Ibid., 112.

[26] Minsky y Papert, *Perceptrons*, 231-32.

[27] Rumelhart y McClelland, *Parallel Distributed Processing*, 112.

[28] Para una evaluación de los éxitos concretos de la estrategia de la representación simbólica hasta 1978, véase Hubert Dreyfus, *What Computers Can't Do*, segunda edición (Nueva York: Harper and Row, 1979).

[29] Newell y Simon, «Computer Science and Empirical Inquiry», 197.

[30] Newell, «Intellectual Issues», 10.

[31] John Haugeland, *Artificial Intelligence: The Very Idea* (Cambridge: MIT Press, 1985) y Margaret Boden, *Artificial Intelligence and Natural Man* (Nueva York, Basic Books, 1977) [Traducción española: *Inteligencia artificial y hombre natural*, Madrid, Tecnos, 1984]. El trabajo sobre redes neuronales continuó marginalmente en psicología y en neurociencia. James Anderson, en la Brown University , siguió defendiendo un modelo de red en psicología, aunque él mismo tenía que vivir de los subsidios de otros investigadores; Stephen Grossberg elaboró una elegante implementación matemática de las capacidades cognitivas elementales. Sobre la postura de Anderson, véase «Neural Models with Cognitive Implications», en *Basic Processing in Reading*, editado por D. LaBerse y S. J. Samuels (Hillsdale, N. J.; Lawrence Erlbaum Associates, 1978). Para ejemplos del trabajo de Grossberg en los años oscuros, véase el libro *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition and Motor Control* (Boston: Reidel Press, 1982). El trabajo inicial de Kohonen se describe en *Associative Memory – A System-Theoretical Approach* (Berlín: Springer-Verlag, 1977).

En el MIT Minsky siguió dictando cátedras sobre redes neuronales y asignando tesis para investigar sus propiedades lógicas. Pero, de acuerdo con Papert, Minski sólo lo hacía porque las redes tenían propiedades matemáticas interesantes, mientras que nada de interés puede probarse respecto de las propiedades de los sistemas de símbolos. Más aún, muchos investigadores en IA asumían que dado que las máquinas de Turing eran manipuladoras de símbolos, y dado que Turing había demostrado que las máquinas de Turing podían demostrar cualquier cosa, él había demostrado que mediante la lógica se podía capturar cualquier cosa inteligible. Según esta perspectiva, una perspectiva holística (y en aquellos días, estadística) necesita justificación, mientras que la estrategia simbólica de la IA no la requiere. Esta confianza, sin embargo, se basaba en una confusión entre los símbolos no interpretados de una máquina de Turing (ceros y unos) con los símbolos semánticamente interpretados de la IA.

[32] Martin Heidegger, *Being and Time* (Nueva York: Harper and Row, 1962), sec. 14-21 [Traducción española: *El Ser y el Tiempo*, México, Fondo de Cultura Económica, 1951]; Hubert Dreyfus, *Being-in-the-World: A Commentary on Division I of Being and Time* (Cambridge: MIT Press, próxima publicación, 1988).

[33] De acuerdo con Heidegger, Aristóteles estaba más cerca de cualquier otro filósofo de comprender la importancia de la actividad cotidiana, pero incluso él sucumbió a la distorsión del fenómeno del mundo cotidiano implícito en el sentido común.

[34] Leibniz, *Selections*, 48.

[35] Terry Winograd, «Artificial Intelligence and Language Comprehension», en *Artificial Intelligence and Language Comprehension* (Washington, D.C.: National Institute of Education, 1976), 9.

[36] Ludwig Wittgenstein, *Philosophical Investigations* (Oxford, Basil Blackwell, 1953) [Traducción española: *Investigaciones filosóficas*, México, UNAM].

[37] Ludwig Wittgenstein, *Philosophical Remarks* (Chicago, University of Chicago Press, 1975).

[38] Wittgenstein, *Philosophical Investigations*, 21.

[39] Véase Husserl, *Intentionality and Cognitive Science*, editado por Hubert Dreyfus (Cambridge: MIT Press, 1982).

[40] «Der Sinn... so wie wir ihn bestimmt haben, ist nicht ein konkretes Wesen im Gesamtbestande des Noema, sondern eine Art ihm einwohnender abstrackter Form.» Véase Edmund Husserl, *Ideen Zu Einer Reinen Phänomenologie und Phänomenologischen Philosophie* (La Haya: Nijhoff, 1950). Para evidencia en el sentido de que Husserl sostenía que el noema daba cuenta de la intencionalidad de la actividad mental, véase Hubert Dreyfus, «Husserl's Perceptual Noema», en *Husserl, Intentionality and Cognitive Science.*

[41] Edmund Husserl, *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy*, traducción de F. Kersten (La Haya: Nijhoff, 1982) [Traducción española de José Gaos, *Ideas Relativas a una Fenomenología Pura y Filosofía Fenomenológica*, México, 1949].

[42] Edmund Husserl, *Cartesian Meditations*, traducción de D. Cairns (La Haya, Nijhoff, 1960), 45.

[43] Ibid., 53.

[44] Ibid., 51.

[45] Marvin Minsky, «A Framework for Representing Knowledge,» en *Mind Design*, editado por John Haugeland (Cambridge: MIT Press, 1981), 96.

[46] Heidegger, 121-122.

[47] Edmund Husserl, *Crisis of European Sciences and Transcendental Phenomenology*, traducción de D. Carr (Evanston: Northwestern University Press, 1970), 149 [Traducción española: *La Crisis de las Ciencias Europeas y la Filosofía Trascendental*, Buenos Aires, 1960].

[48] Edmund Husserl, *Ideen zu einer reinen Phänomenologie und Phenomenologischen Philosophie*, libro 3 del vol. 5, *Husserliana* (La Haya: Nijhoff, 1952), 134.

[49] Husserl, *Cartesian Meditations*, 54-55.

[50] Husserl, *Formal and Trascendental Logic*, traducción de D. Cairns (La Haya, Nijhoff, 1969), 244 [Traducción española de L. Villoro: *Lógica Formal y Lógica Trascendental*, México, 1962].

[51] Ibid., 246.

[52] Husserl, *Crisis*, 291.

[53] Minsky, «A Framework», 124.

[54] Newell, «Intellectual Issues», 222-223.

[55] Herbert Simon, *The Shape of Automation for Men and Management* (Nueva York, Harper and Row, 1965), 96.

[56] Terry Winograd, «Computer Software for Working with Language», *Scientific American* (setiembre de 1984):142.

[57] Esta concepción de la habilidad se explica y se defiende en Hubert Dreyfus y Stuart Dreyfus, *Mind Over Machine* (Nueva York: Macmillan, 1986).

[58] Imre Lakatos, *Philosophical Papers*, editado por J. Worrall (Cambridge: Cambridge University Press, 1978).

[59] Marvin Minsky, *Computation: Finite and Infinite Machines* (Nueva York: Prentice-Hall, 1977), 2.

[60] Gina Kolata, «How can Computers get Common Sense?», *Science 217* (24 de setiembre de 1982):1237.

[61] Pamela McCorduck, *Machines Who Think* (San Francisco: W.H. Freeman, 1979), 87.

[62] Citas típicas de *Principles of Neurodynamics* de Rosenblatt:

"En un experimento de aprendizaje, un perceptrón está típicamente expuesto a una secuencia de patrones que contienen elementos representativos de cada tipo o clase que ha de distinguirse, y la elección apropiada se «refuerza» conforme a alguna regla de modificación de la memoria. Se presenta entonces al perceptrón un estímulo de prueba, y se determina la probabilidad de que dé la respuesta apropiada para esa clase de estímulo. Si el estímulo de prueba activa un conjunto de elementos sensoriales que son enteramente distintos de los que fueron activados en exposiciones previas a estímulos de la misma clase, el experimento es una prueba de «generalización pura». El más simple de los perceptrones no tiene capacidad de generalización, pero se puede mostrar que se comporta muy respetablemente en experimentos de discriminación, en particular si el estímulo de prueba es casi idéntico a alguno de los patrones previamente experimentados." (p. 68)

"Se considera hoy que los perceptrones son muy poco parecidos a los sujetos humanos en su habilidad para detectar figuras y en sus tendencias a organizar totalidades gestálticas." (p. 71).

"El reconocimiento de secuencias en formas rudimentarias se encuentra plenamente dentro de las capacidades de los perceptrones bien organizados, pero el problema de la organización figural y de la segmentación presenta problemas que son tan serios aquí como en el caso de la percepción estática de patrones." (p. 72)

"En un perceptrón simple, los patrones se reconocen antes que las «relaciones»; por cierto, las relaciones abstractas, tales como «A se encuentra encima de B» o «el triángulo está dentro del círculo» nunca se abstraen como tales, sino que se pueden adquirir por medio de una especie de procedimiento exhaustivo de aprendizaje de memoria, en el que se enseña individualmente al perceptrón cada uno de los casos en los que se mantiene la relación." (p. 73)

"Una red consistente en menos de tres niveles de unidades de trasmisión de señales, o una red que consista sólo de elementos lineales conectados en serie, es incapaz de aprender a discriminar clases de patrones en un ambiente isotrópico (donde puede presentarse cualquier patrón en todas las posiciones retinianas posibles, sin efectos de borde)." (p. 575)

"En los capítulos precedentes hemos presentado cierto número de modelos especulativos que aparentemente son capaces de aprender programas secuenciales, fragmentar el discurso en fonemas y aprender los «significados» sustantivos de nombres y verbos mediante referentes sensoriales simples. Esos sistemas representan los límites superiores en la conducta abstracta en los perceptrones según se cree hasta hoy. Los perceptrones se encuentran limitados por la falta de una «memoria temporaria» satisfactoria, por su falta de habilidad para percibir relaciones topológicas abstractas en forma simple y de aislar entidades figurales significativas, u objetos, salvo bajo condiciones especiales." (p. 577)

"Las aplicaciones más susceptibles de realizarse con los tipos de perceptrones descritos en este volumen incluyen reconocimiento de caracteres y «máquinas de

leer», reconocimiento del habla (de palabras distintas, claramente separadas) y capacidades extremadamente limitadas de reconocimiento pictórico, o reconocimiento de objetos contra un fondo simple. La «percepción», en un sentido amplio, puede estar potencialmente dentro de la capacidad de los descendientes de nuestros modelos actuales, pero debe obtenerse mucho más conocimiento fundamental antes que se pueda prescribir un diseño lo suficientemente sofisticado como para que un perceptrón pueda competir con el hombre bajo condiciones ambientales normales." (p. 583)

[63] Rumelhart y McClelland, *Parallel Distributed Processing*, vol. 1, 45.

[64] Ibid., vol. 2, 535.

[65] Paul Smolensky, «On the Proper Treatment of Connectionism», *Behavioral and Brain Sciences*, próximo a aparecer.

[66] Geoffrey Hinton, «Learning Distributed Representations of Concepts», en *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (Amherst, Massachusetts: Cognitive Science Society, agosto de 1986).

[67] Para una reseña reciente de la percepción en la que se niega la necesidad de representaciones mentales, véase James J. Gibson, *The Ecological Approach to Visual Perception* (Boston: Houghton Mifflin, 1979). Gibson y Rosenblatt colaboraron en un ensayo de investigación para la Fuerza Aérea de los Estados Unidos en 1955; véase J. J. Gibson, P. Olum y F. Rosenblatt, «Parallax and Perspective During Aircraft Landing», *American Journal of Psychology* 68 (1955):372-85.

3
Inteligencia natural e inteligencia artificial

Robert Sokolowski

En este ensayo no intentaremos decidir si la inteligencia artificial es lo mismo que la inteligencia natural. En lugar de eso, examinaremos algunas de las cuestiones y términos que se deben poner en claro antes que esta pregunta pueda contestarse. Discutiremos de qué forma puede formularse la pregunta sobre la relación entre la inteligencia natural y la artificial.

Una de las primeras cosas que hay que aclarar es la ambigua palabra *artificial*. El adjetivo se puede usar en dos sentidos, y es importante saber cuál se aplica en el término *inteligencia artificial*. La palabra *artificial* se utiliza en un sentido cuando se aplica, digamos, a flores, y en otro sentido cuando se aplica a la luz. En ambos casos, algo se llama artificial porque es fabricado. Pero en el primer uso artificial significa que la cosa parece ser, pero realmente no es, lo que parece. Lo artificial es lo meramente aparente; sólo muestra cómo aparenta ser otra cosa. Las flores artificiales son sólo papel, no flores en absoluto; cualquiera que las tome por flores está equivocado. Pero la luz artificial es luz e ilumina. Es fabricada como un sustituto de la luz natural, pero una vez fabricada es lo que parece ser. En este sentido lo artificial no es lo meramente aparente, ni simplemente una imitación de otra cosa. La apariencia de la cosa revela lo que es, no cómo parece ser otra cosa.

El movimiento de un automóvil es otro ejemplo de algo que es artificial en el segundo sentido de la palabra; un automóvil se mueve artificialmente; se mueve sólo porque los seres humanos lo han construido para moverse y para hacerlo andar liberando energía almacenada. Pero se mueve realmente, no sólo parece estar movién-

---

Robert Sokolowski. Profesor de filosofía en la Universidad Católica de los Estados Unidos.

dose. En constraste, el panel de madera artificial del automóvil sólo parece ser de madera; arde, se dobla, se quiebra y decae como plástico, no como madera. Parece ser de madera sólo a la visión y sólo desde cierto ángulo y con ciertas clases de luz.

¿En qué sentido usamos la palabra *artificial* cuando hablamos de inteligencia artificial? Los críticos de la inteligencia artificial, los que calumnian la idea y dicen que ha sido inflada y sobrevendida, afirmarían que el término se usa en el primer sentido, para significar lo meramente aparente. Dirían que la inteligencia artificial no es en realidad nada más que complejas estructuras mecánicas y procesos eléctricos que presentan (a los crédulos) una ilusión de cierta clase de pensamiento. Los que apoyan la idea de la inteligencia artificial, los que afirman que el término denota algo genuino y no meramente aparente, dirían que la palabra *artificial* se usa en el segundo de los sentidos que hemos distinguido. Obviamente, dirían, las máquinas que piensan son artefactos; obviamente son ejecutadas por seres humanos; pero una vez que están en movimiento, las máquinas piensan. Su pensamiento puede ser diferente del de los seres humanos en algunos respectos, así como el movimiento de un automóvil es diferente del de un conejo y el vuelo de un avión es diferente del de un ave, pero es una especie de pensamiento genuino, así como es genuino el movimiento de un automóvil y genuino el vuelo de un avión.

Supongamos que afirmamos que la inteligencia artificial es una inteligencia genuina, aunque construida. ¿Debemos probar la verdad de esa afirmación? ¿Estamos obligados a demostrar que las máquinas piensan realmente, que no sólo parecen tener inteligencia? Quizá no; nadie ha probado que la luz artificial ilumina y que los aviones realmente vuelan. Solamente vemos que lo hacen. Si las máquinas que piensan exhiben la actividad de pensar, ¿por qué no debemos admitir que son realmente inteligentes?

El problema es que el pensamiento no es tan visible y palpable como lo son la iluminación, el movimiento y el vuelo; no es tan fácil decir si el pensamiento está presente o no. Aun cuando hablamos con otro ser humano, no siempre podemos estar seguros si la persona está hablando y actuando pensadamente o meramente recitando de memoria, comportándose automáticamente. Y hay casos en que las máquinas sólo parecen pensar pero realmente no lo hacen: la calculadora electrónica puede realizar cosas notables, pero sólo alguien que haya sido engañado por ellas (alguien como

quien toma las flores artificiales por flores reales) diría que la calculadora posee una inteligencia propia. La calculadora puede revelar la inteligencia de quien la construyó y la programó, pero ella no origina su propio pensamiento.

¿En qué difiere la inteligencia artificial de la calculadora? ¿En qué es diferente de la computación numérica? ¿Qué es lo que hace que podamos llamar el pensamiento propio de una máquina, una actividad propia que no pueda ser disuelta en la inteligencia de la gente que construyó y programó la máquina? Si fuéramos a afirmar que la máquina pensante, aunque sea un artefacto, exhibe inteligencia, deberíamos aclarar qué queremos significar por el «pensamiento» que se dice que ella ejecuta. Esta puede no ser una prueba, pero es una explicación, y parece ser que se requiere demasiada justificación para sustentar nuestra afirmación de que las máquinas piensan.

Alan Turing asentó el principio de que si una máquina se comporta inteligentemente, debemos acreditarle inteligencia.[1] La conducta es la clave. Pero la prueba de Turing no puede seguir siendo por sí sola el criterio para dirimir la inteligencia de las máquinas. El pensamiento de máquina sólo reproducirá parte del pensamiento natural; podrá limitarse, por ejemplo, a respuestas producidas sobre una pantalla. A este respecto, nuestra experiencia del pensamiento de la máquina es como hablar con alguien por teléfono, y no como estar con otra persona y verla actuar, hablar y responder a nuevas situaciones. ¿Cómo sabemos que nuestra visión parcial de la inteligencia de la máquina no es como ese ángulo de visión desde el que las flores artificiales nos parecen reales? ¿Cómo podemos saber que no estamos siendo engañados si estamos atrapados en una perspectiva desde la cual una inteligencia meramente aparente se parece mucho a la inteligencia real? Hay que agregar a la prueba de Turing alguna clase de argumentación para mostrar que la inteligencia artificial es artificial en el segundo sentido de la palabra y no en el primero, que aunque sea construida y parcial, sigue siendo genuina y no sólo aparente. Necesitamos decir más sobre la inteligencia para mostrar si es real o no, y necesitamos poner en claro la diferencia entre sus formas naturales y artificiales.

I

Al discutir la distinción entre inteligencia natural y artificial, debemos guardarnos de establecer divisiones abruptas e ingenuas. Si formulamos nuestra pregunta en términos de alternativas absolutas, podemos poner nuestra argumentación en una camisa de fuerza y privarla de la flexibilidad que necesita. Con esta aproximación rígida opondríamos la computadora al cerebro, considerando la inteligencia natural como una actividad desarrollada por el cerebro y la inteligencia artificial como una actividad desarrollada en las computadoras. Aquí el cerebro, allá la computadora; aquí la inteligencia natural, allá la inteligencia artificial. La actividad es definida por el material en el que se manifiesta.

Esta aproximación es poco aguda y niega algo que vincula la inteligencia natural y la artificial: la palabra escrita. La inteligencia artificial no imita simplemente el cerebro y el sistema nervioso; ella transforma, codifica y manipula discurso escrito. Y la inteligencia natural no es sólo una actividad orgánica que ocurre en un cerebro en funcionamiento; también está incorporada en las palabras escritas en un papel, inscritas en arcilla, pintadas sobre un cartel. Entre el cerebro y la computadora está la escritura.

Cuando el pensamiento se corporiza en la palabra escrita, hay algo artificial en él. Consideremos un signo de neón parpadeante que dice «Hotel». Las personas no reaccionan a ese signo como lo harían ante una piedra o un árbol. Leen el signo y también lo contestan. Se comportan ante él de una manera análoga a la que tendrían ante alguien que les dijera que el edificio es un hotel y que pueden conseguir un cuarto allí. Más aún, la persona que puso el signo donde está (la persona que afirma algo mediante el signo y la que es responsable por decir lo que el signo dice) no tiene que permanecer cerca del signo para obtener su efecto. Puede dejar que el signo siga solo; éste funciona sin él. Es un artificio, un artificio que manifiesta y comunica algo a alguien, invitando tanto a una interpretación como a una respuesta.

Por supuesto, la inteligencia artificial promete hacer algo más que lo que ya puede hacer la escritura, pero tiene un pie en ella: pone en movimiento el pensamiento que está corporizado en la escritura. Nuestro desafío filosófico es poner en claro qué tipo de movimiento es el pensamiento.[2] La continuidad entre la escritura y la inteligencia artificial nos mostrará menos aprehensivos sobre el hecho de ser reemplazados de algún modo por máquinas pensantes. En cierta

forma, ya estamos siendo reemplazados por el mundo de la escritura. Si dejo instrucciones escritas detrás de mí, no tengo que estar presente para que las instrucciones tengan efecto. Pero esto no cancela mi pensamiento; lo enriquece. Si encontramos registros escritos en las ruinas de una antigua ciudad, no pensamos que los hablantes de esa ciudad quedaron anulados como hablantes por los documentos, o que éstos destruyeron su subjetividad; pensamos que su discurso era apreciado más vívidamente como discurso en contraste con la palabra escrita. También creemos que su pensamiento ha sido amplificado por su escritura, no suprimido por ella, porque a través de la palabra escrita han sido capaces de «hablar» con nosotros. De igual modo, la codificación de la escritura en la inteligencia artificial no significa que no tengamos que pensar más. Más bien, nuestro propio pensamiento puede ser apreciado más vívidamente en contraste con el que pueden hacer las máquinas; el hecho de que algunas dimensiones del pensamiento puedan ser llevadas a cabo mecánicamente nos pone más vívidamente de manifiesto las que sólo nosotros podemos ejecutar. Si el pensamiento artificial puede sustituir algunos de nuestros pensamientos como la luz artificial sustituye a la luz natural, las formas de pensamiento para las que ninguna sustitución sea posible resaltarán más claramente como las nuestras propias.

La difusión gradual de la escritura en los asuntos humanos puede servir como un análogo histórico de la corriente de aire de la inteligencia artificial en los intercambios humanos. La escritura no sólo reemplazó las actividades lingüísticas que desarrollaba la gente antes que hubiera escritura; su principal impacto fue el de hacer posibles nuevas clases de actividades y darles una nueva forma a las viejas. Amplió y diferenció las actividades económicas, legales, políticas y estéticas, e hizo a la historia posible. Incluso permitió que la religión tuviera una forma nueva: permitió el surgimiento de religiones que involucraban libros, con todas las cuestiones concomitantes de texto, interpretación y comentario. La escritura hizo todo esto ampliando la inteligencia. La imprenta aceleró la difusión de la palabra escrita, pero no cambió la naturaleza de la escritura.

La pregunta que se puede plantear ante la inteligencia artificial es si ella es meramente una extensión de la escritura o un reajuste en las empresas humanas que comenzó cuando la escritura penetró en los asuntos humanos. El procesamiento de textos es claramente un refinamiento de la imprenta, una especie de tipografía glorificada, pero la inteligencia artificial parece ser más que eso. Parece ser

capaz de reformular la corporización del pensamiento que se había logrado en y por la escritura. ¿Qué resultará ser la inteligencia artificial? ¿Será un poscrito a la escritura, o la escritura resultará ser un preludio de cuatro mil años a la inteligencia artificial? El impacto de la escritura ¿radicará en haber sido una preparación para el pensamiento mecánico?

Si la inteligencia artificial es por cierto una transformación de la escritura, entonces es más parecida a la luz artificial y menos parecida a las flores artificiales: un sustituto genuino para ciertas clases de pensamiento, no sólo una imitación superficial. El pensamiento es conformado por la escritura; la inteligencia se modifica cuando toma forma escrita; la escritura nos permite identificar y diferenciar cosas en formas que no eran posibles cuando podíamos hablar pero no escribir. Si la inteligencia artificial puede a su vez transformar la escritura, podrá encarnar una clase de inteligencia que no existiría de otra manera, lo mismo que el automóvil nos proporciona una clase de movimiento que no estaba disponible antes que fuera inventado.

En el caso de cualquier nueva tecnología, lo nuevo se comprende primero dentro del horizonte definido por lo viejo. Los primeros automóviles, por ejemplo, se parecían mucho a los carruajes. Toma tiempo para que las nuevas posibilidades se afirmen por sí mismas, para que tomen forma tanto ellas como el ambiente en el que deben encontrar su lugar. Le tomó tiempo al automóvil generar autopistas y garajes. Los sistemas expertos diseñados en las primeras etapas de la inteligencia artificial están siguiendo este patrón.[3] Pretenden reemplazar una forma de pensamiento más bien prosaica, una forma que ya parece madura para el reemplazo: la clase de pensamiento ejercida por la oficina de informes o por el farmacéutico, una persona que conoce una cantidad de hechos y que puede coordinarlos y trazar algunas de sus implicaciones. Los sistemas expertos son los carruajes sin caballos de la inteligencia artificial. Son análogos a las escrituras antiguas que solamente registraban los contenidos del tesoro real o la distribución del grano.

Esto no va en detrimento de los sistemas expertos. Los reemplazos iniciales, pequeños y obvios de las viejas maneras de hacer las cosas deben asentarse antes que puedan tener lugar los logros más distintivos de una nueva forma intelectual, en este caso, antes que puedan surgir los Dantes, Shakespeares y Newtons, o los Jaguars, las autopistas y las estaciones de servicio de la inteligencia artificial. Y así como los que experimentaron los comienzos de la escritura

difícilmente pudieran imaginar lo que Borges y Bohr podrían hacer, o lo que podría ser una biblioteca nacional, o un centro de investigación médica, o un contrato de seguros, igualmente nosotros (si la inteligencia artificial es ciertamente una renovación de la escritura) dudosamente podríamos concebir qué formas podrá tomar el florecimiento del pensamiento de máquina.

Más aún, gran cantidad de pensamiento humano es más bien mecánico. Sólo demanda que estemos bien informados y que seamos capaces de registrar relaciones y trazar inferencias dentro de lo que conocemos. La medida en que este pensamiento de rutina penetra nuestra actividad intelectual sólo se advierte cuando el pensamiento artificial logra realizar este trabajo por nosotros.[4] Enormes áreas de la compilación de datos científicos, mediciones y correlaciones, del planeamiento de estrategias en tasaciones o seguros, de la elaboración de combinaciones aceptables de antibióticos y de sus correspondencias con infecciones, del diseño de redes y planes para el tráfico aéreo, de la resolución de formas de tratar con leyes y regulaciones, son tareas que se pueden codificar y organizar de acuerdo con reglas especificables. La inteligencia artificial será rápidamente capaz de liberarnos de estos trabajos laboriosos. Pero, dado que existen pocas bendiciones sin mezcla, es probable que también introduzca nuevas rutinas, faenas penosas y complejidades nada bienvenidas que no hubieran surgido si no hubiesen existido las computadoras.

Estamos comprensiblemente asombrados de la forma en que las máquinas pueden almacenar conocimiento e información, y cómo incluso parecen «pensar» con ese conocimiento e información. Pero estas capacidades de las máquinas no deben impedirnos ver algo que es más simple pero quizás aún más extraño: el pavoroso almacenamiento y representación que tiene lugar cuando el conocimiento toma cuerpo en la palabra escrita. En la inteligencia artificial el cuerpo cambia, pero la diferencia principal es el nuevo tipo de material corporal, no el cuerpo en sí. La luz de neón que parpadea la palabra Hotel retiene muchos de los rasgos que se encuentran en las máquinas pensantes: hay un significado disponible, se indica un curso de comportamiento, se legitiman inferencias. Parece no haber nadie que hable o posea los significados —el significado parece flotar— y sin embargo hay algún significado en el signo. El significado está disponible para todos y parece sobrevivir a cualquier hablante individual.

En inteligencia artificial tales significados toman cuerpo en materiales que permiten manipulaciones extremadamente complejas de un tipo sintáctico. De aquí que la máquina parezca razonar, mientras que el signo no parece razonar sino sólo afirmar. En lugar de comparar meramente computadoras y cerebros, deberíamos asimismo comparar el «razonamiento» de la máquina con la «afirmación» del signo, y examinar el razonamiento y la representación a medida que ocurren en la máquina y en la escritura.

Es cierto que la inteligencia artificial puede ir más allá de las impresiones hacia un discurso artificialmente hablado. Si logra hacerlo, su «habla» habrá sido una transformación de la escritura y llevará la impronta de la escritura. La inteligencia artificial se habrá movido en una dirección que es la inversa a la que siguió el pensamiento natural, que se movió de la palabra hablada a la palabra escrita.

II

La palabra escrita puede servir como agente entre la inteligencia natural y la artificial. Está a caballo de ambas: la inteligencia natural toma cuerpo y se modifica en la escritura, pero la escritura es de algún modo artificial, algo así como un artefacto. Investiguemos más de cerca este papel mediador de la escritura. ¿Cómo puede servir la escritura de puente hacia la inteligencia artificial?

Rondaremos en torno de esta cuestión haciendo una pregunta más general sobre las condiciones necesarias para el surgimiento de la inteligencia artificial: ¿Qué se requiere para que la inteligencia artificial llegue a existir? Una respuesta obvia es que se necesitan ciertos lenguajes de computación, como LISP y Prolog. Otra es que se requieren las computadoras mismas, con el hardware, la arquitectura y la memoria adecuada. Otra respuesta más es que la lógica matemática desarrollada en los pasados cien años o algo así por Gottlob Frege, Giuseppe Peano, Bertrand Russell y otros fue una condición necesaria tanto para el hardware como para el software que hoy tenemos. Resulta interesante que estos avances en matemáticas y lógica se hayan desarrollado por razones puramente teoréticas, para demostrar, por ejemplo, que la aritmética es una parte de la lógica pura (el objetivo de Frege) y no para preparar un lenguaje para máquinas pensantes.[5] La aplicación tecnológica tomó

ventaja de la apertura proporcionada por el logro teorético: «La oportunidad creó el apetito, la oferta creó la demanda»[6].

Todos estos prerrequisitos para la inteligencia artificial —el hardware de computadora, los lenguajes de computadora y las lógicas formalizadas— fueron logrados por personas identificables en momentos definidos. Podemos dar los nombres y las fechas de su invención. Pero hay otra condición habilitante que es de otra naturaleza. Es de un interés filosófico mucho más grande, y es también mucho más elusiva; es difícil decir cuándo apareció en escena y quién fue el responsable por haberla traído. Pero sin ella, no podrían haber surgido ni la inteligencia artificial ni ninguno de sus prerrequisitos. La condición de marras es que podemos abordar un signo lingüístico de dos maneras. Podemos pensar qué es lo que el signo expresa o podemos pensar acerca del signo y de la forma en que está compuesto.

Para ilustrar esta diferencia, consideremos dos formas de traducción.[7] Consideremos primero al traductor que trabaja en encuentros internacionales y que traduce discursos a medida que se producen. Ese traductor piensa sobre el tópico que se discute. Si el discurso es sobre barcos oceánicos, el traductor piensa en barcos, carga, leyes, líneas de costa y corrientes marinas. Habla acompañando al hablante original; puede anticipar algunas de las frases y palabras del hablante y puede incluso hablar antes que él. El traductor puede hacerlo porque está guiado por las cosas de las que se habla y que se presentan en el discurso; no se concentra en el análisis de las palabras.

En contraste, consideremos el caso de alguien que está aprendiendo griego y tratando de traducir un texto en esa lengua. Inspecciona cada palabra, advierte los sufijos, establece cuál palabra puede ser un verbo y cuál un sujeto, trata de imaginarse la forma en que una palabra puede ser el resultado de ciertas elisiones y contracciones, procura determinar qué significa y cómo encaja con otras palabras. Gradualmente va definiendo un sentido posible para la frase. En este caso lo que se expresa en las palabras no orienta la traducción; más bien, la cosa significada viene a lo último, sólo después que las palabras han sido el centro de atención durante cierto tiempo. Este traductor no puede anticipar las palabras del hablante porque no está siendo guiado por los temas que se discuten. En este caso, las cosas expresadas están al margen de la atención mientras las palabras están en el centro, en tanto que en el primer caso las palabras están en los márgenes mientras el tema expresado está en el foco.

Podemos pasar del foco en lo que se expresa a un foco en las palabras, podemos ir de vuelta a las cosas expresadas y podemos ir y volver una y otra vez. Cuando estamos en uno de los focos, el otro siempre permanece en un margen como un foco al que podemos entrar. Y los dos focos no son meramente anexos mutuos; cada uno es lo que es en conjunción con el otro. Concentrarse en las palabras como palabras es posible porque se ejecuta contra el foco de lo que las palabras expresan; el foco en los temas es lo que es (para nosotros como hablantes) sólo cuando se lo ejecuta contra el foco del que trata o pudiera tratar el asunto. No hay palabras, excepto las que se definen en esta doble perspectiva; no hay palabras que estén meramente «alrededor».

Ahora bien, la inteligencia artificial es posible porque podemos volver nuestra atención decididamente hacia la palabra y, en lugar de analizar su composición gramatical o fonológica, podemos comenzar a codificar la palabra, reemplazar sus letras por una serie de dígitos binarios y sus posibilidades sintácticas por operaciones computadorizadas. Podemos alfabetizar y gramatizar la palabra de una nueva forma. Podemos reducirla a secuencias de unos y ceros y a reglas de manipulación. Pero al hacerlo nunca cancelamos nuestra apreciación de que estamos tratando con una palabra, y que ella expresa cierta cosa; nunca arrancamos el margen en el que se expresa el certificado. Por esta razón el resultado final de nuestra codificación y transformación continúa expresando algo. Por esta razón llamamos al resultado de lo que hacemos nosotros y las máquinas una *inteligencia* artificial, una comprensión de algo, no meramente una nueva disposición de marcas.

Es aquí donde se debe explorar y comprender la «intencionalidad» de los programas de computadora: no preguntando si una computadora es igual a un cerebro, sino preguntando de qué modo los eductos de la computadora son como palabras escritas, y cómo este cambio de foco, de pensar en las expresiones a pensar en lo que se expresa, puede todavía tener lugar con respecto al «habla» que los procesos que se desenvuelven en la máquina pensante liberan para nosotros.[8]

Puede incluso resultar ambiguo decir que la palabra ha de tener un significado, porque el significado puede aparecer entonces como una entidad de alguna clase que se sitúa entre la palabra y la cosa que ella representa. En algunas teorías sobre la cognición, tal significado sustancializado se localiza en el cerebro o en la mente;

pudiera seguirse de ello una discusión sobre si este significado también ha de encontrarse como alguna especie de representación en la computadora y sus programas.[9] Pero no hay necesidad de tal entidad. Todo lo que necesitamos hacer es admitir la capacidad que está en nosotros de concentrarnos sobre la palabra mientras la cosa que representa queda al margen, o focalizar la cosa mientras queda al margen la palabra que la simboliza. No se necesita nada más. El significado es simplemente la cosa en tanto significada por la palabra.

Sabemos que Frege desarrolló su nueva notación lógica en los años anteriores a 1879, cuando se publicó su *Begriffsschrift*. Pero ¿cuándo se dio cuenta alguien de que podemos concentrarnos en las palabras como palabras y que podemos tomarlas aparte, incluso teniendo en mente lo que ellas significan? No hay fecha para esto; se remonta a mucho tiempo atrás. ¿Y quién es el alguien que apreció esto? La inteligencia artificial está fuertemente endeudada con él, y también lo estamos todos, dado que difícilmente podríamos imaginarnos sin esta habilidad. Podemos volvernos hacia la palabra incluso si estuviéramos limitados a la palabra hablada, pero podemos volvernos hacia ella mucho más explícita, decisiva y analíticamente una vez que hemos comenzado a escribir. Pero la habilidad hacia la que debemos poner el foco precede a la escritura y hace a la escritura posible, y también precede a y permite la ulterior codificación de la escritura que se manifiesta en la inteligencia artificial.

Nuestra habilidad para llevar la atención de, digamos, un árbol a la palabra árbol, nos ayuda a explicar la forma en que las palabras se establecen como símbolos y la forma en que las cosas quedan nombradas por palabras o significadas mediante símbolos. Pero esta habilidad para apartar la atención también puede ayudarnos a aproximarnos a uno de los más debatidos problemas asociados con la inteligencia artificial y la inteligencia natural: el problema de la forma en que los sucesos fisiológicos en el cerebro pueden representar algo que ocurre en el mundo, el problema de cómo describir representaciones mentales, imágenes mentales y símbolos mentales. Las redes neuronales se activan.¿Cómo es que esas activaciones son más que sólo una tormenta eléctrica o un proceso químico en el cerebro? ¿Cómo es que sirven para representar algo que está más allá de ellas mismas y más allá del cerebro? ¿Cómo es que sirven para presentar esta lámpara? ¿Cómo hemos de describir la «palabra del cerebro» o la «imagen del cerebro» de la lámpara?

La mayoría de los escritores que discutió esta cuestión dice simplemente que hay un símbolo mental o representación que hace el trabajo, pero esto no diferencia el «símbolo del cerebro» de otras clases de símbolos con las que normalmente tratamos: los que encontramos en el sonido, en el papel, en lienzos, en madera, en piedra. Una diferencia crucial entre el símbolo del cerebro y el símbolo normal, «público», parece ser la siguiente. En el caso del símbolo público, podemos situar el foco ya sea en el símbolo o en la cosa que él significa: en *esta lámpara* o en la lámpara misma. Pero en el caso de un símbolo del cerebro, el individuo no puede concentrarse sobre la activación neuronal de su cerebro; sólo puede atender al objeto presentado, la lámpara. El símbolo del cerebro es esencial y necesariamente transparente para el individuo. Pero quien se concentra en el mundo cerebral (el neurólogo, digamos, quien examina las activaciones neuronales involucradas en el hecho de ver esta lámpara) no puede ver estas activaciones como una presentación de la lámpara; no puede significar la lámpara a través de ellas (como podría marginalmente significarla mientras se concentra en el término la *lámpara*). Para el neurólogo, las activaciones cerebrales son esencial y necesariamente opacas; son un fenómeno biológico por derecho propio. Para él no son simbólicas, ni siquiera marginalmente. La persona a quien examina debe decirle que es una lámpara lo que está viendo.

De este modo, la palabra cerebral no es como la palabra hablada, pero la reflexión sobre la forma en que constituimos la palabra hablada no puede ayudar a aclarar la naturaleza intrigante de la palabra cerebral. Una persona, en el caso del símbolo público, puede alternar entre el símbolo y la cosa; pero en el caso del símbolo cerebral, éste está fragmentado en dos personas: quien ve la cosa pero no el símbolo cerebral y quien ve el suceso cerebral pero no la cosa. Estas afirmaciones, desde luego, son sólo el comienzo de un análisis de las representaciones mentales, pero indican que una de las mejores maneras de adaptar nuestro lenguaje para describir adecuadamente el cerebro es contrastar el símbolo cerebral con el símbolo público y elaborar este contraste en todos sus detalles. Antes en este ensayo utilizamos la corporización del significado que se manifiesta en la escritura como una ayuda para describir la inteligencia artificial; aquí utilizamos la corporización del significado que se encuentra en los símbolos públicos como una ayuda para comprender la representación que se manifiesta en la inteligencia natural.

Dejemos ahora la cuestión de la representación mental y volvamos a la palabra escrita. Hemos dado por sentado que la escritura de marras es la escritura alfabética, la clase de escritura familiar a los hablantes del inglés. Pero también existe una escritura ideogramática, y será interesante comparar la escritura alfabética con la ideogramática con referencia al cambio de foco que hemos descrito, el cambio que va de prestar atención a la cosa a prestar atención a la palabra.

Puesto que es algo que se parece a una pintura de la cosa representada, una ideografía mantiene esa cosa vívida en la mente aun cuando uno se vuelva hacia la palabra escrita.[10] Una palabra alfabética, por otro lado, abandona toda imagen del objeto y simboliza los sonidos de la palabra hablada. La escritura ideogramática nos empuja hacia la cosa, la escritura alfabética nos empuja a la palabra, pero ninguna puede cortar el otro de sus dos focos. No serían escritura si lo hicieran.

La inteligencia artificial ha trabajado primariamente con simbolismo alfabético. Es interesante especular si algunos rasgos de la escritura ideogramática podrían encontrar un lugar en la inteligencia artificial para complementar, de algún modo, la escritura alfabética. La escritura ideogramática desecha las inflexiones y trae hacia la superficie la estructura gramatical profunda de las frases[11]; estas cualidades podrían simplificar la gramática y la lógica de la narración y hacer que las narrativas sean más fáciles de codificar. La expresión ideogramatical podría no ser útil para crear lenguajes de programación, pero sería útil para modificar lo que los programas presentan para que los usuarios de computadoras lean e interpreten. Una influencia ideogramática en el lenguaje de la interface entre el usuario y la máquina haría que este lenguaje fuera diferente del que normalmente hablamos (produciría una especie de inglés pidgin, por ejemplo) pero estaríamos esperando eso.[12] Nuestro lenguaje natural se ha desarrollado bastante lejos de cualquier compromiso con máquinas pensantes y no está adaptado a ellas. Ha servido a otros propósitos en otras circunstancias. ¿Por qué forzar la inteligencia artificial en las coacciones que se requerirían para que su conducta se pareciera al habla inglesa ordinaria? La máquina pensante es una nueva presencia, como la escritura alguna vez lo fue. Nuestro lenguaje natural, con su adaptabilidad exuberante, encontrará modos de meterse en esto, aunque deba estirarse más allá de su forma alfabética para hacerlo.

III

El tipo de pensamiento que la inteligencia artificial se supone capaz de emular es el razonamiento deductivo inferencial: sacar conclusiones una vez que se han sentado axiomas y reglas de derivación. Hacer deducciones significa alcanzar nuevas verdades sobre la base de las que ya conocemos. Es esta clase de razonamiento lo que Frege trataba de formalizar en su nueva notación lógica, el antepasado de los lenguajes de computadora. Frege quería asegurar la adecuación de las deducciones haciendo explícito y formalmente justificado cada paso de la deducción y manteniendo las derivaciones libres de premisas ocultas. Se suponía que su notación haría posible esa pureza de razonamiento.[13] El producto subsecuente de los esfuerzos de Frege ha sido la lógica y los programas que hacen que las deducciones sean tan explícitas que pueden ser realizadas mecánicamente; por cierto, la parte de un programa de inteligencia artificial que deriva conclusiones se llama a veces con el pintoresco nombre de «máquina de inferencia».

Pero derivar inferencias no es la única clase de inteligencia; hay también otras clases. Discutiremos las citas y las distinciones como dos formas de actividad intelectual que no son reducibles a la derivación de inferencias. También discutiremos el deseo que nos mueve a pensar. Estas formas y aspectos de la inteligencia natural —citar, distinguir, desear— son de interés para la inteligencia artificial de dos maneras. Si la inteligencia artificial puede de alguna manera darles cuerpo, demostrará ser exitosa en el reemplazo de la inteligencia natural. Pero si llega a ser evidente que la inteligencia artificial no puede imitar estos poderes y actividades, habremos descubierto algunos de los límites de la inteligencia artificial y comprenderemos mejor la diferencia entre ambas clases de inteligencia.

La inteligencia artificial depende tanto de la ingeniería como de la fenomenología. La ingeniería es el desarrollo del hardware y de los programas; la fenomenología es el análisis del conocimiento natural, la descripción de las formas de pensamiento que la ingeniería puede o bien tratar de imitar y reemplazar, o tratar de complementarla si no puede hacerlo. Esta discusión es una contribución a la fenomenología de la inteligencia artificial, desarrollada en el contexto definido por los propósitos y las posibilidades de la inteligencia artificial.

Citas

Una de las características esenciales de la inteligencia natural es que como hablantes podemos citarnos mutuamente. Esto no sólo significa que podemos repetir las palabras que algún otro ha dicho; significa que podemos apreciar y establecer cómo parecen ser las cosas para alguien. Nuestra cita de las palabras de otro es meramente la forma en que nos presentamos y presentamos a los demás la forma en que el mundo se aparece para alguien diferente de nosotros mismos.[14] La habilidad para citar nos permite agregar perspectivas a las cosas que experimentamos y expresamos. Veo las cosas no sólo desde mi propio punto de vista, sino como le parecen a otro desde otra perspectiva, a alguien que tiene una historia diferente de la mía, a alguien cuyos intereses difieren de los míos. Es señal de mayor inteligencia ser capaz de apreciar las cosas tal como otros las experimentan, y es señal de menor inteligencia ser incapaz de hacerlo: somos obtusos si sólo vemos las cosas de una manera, nuestra propia manera.

No describimos adecuadamente esta habilidad si la llamamos la capacidad de ponernos en lugar de otro, como si la cosa importante fuera compartir los estados de ánimo y los sentimientos de esa persona, simpatizar con sus estados subjetivos. Incluso los estados de ánimo y los sentimientos que buscamos compartir son una respuesta a la forma en que parecen ser las cosas, y la forma en que las cosas son para alguien se puede capturar en una cita. Además, puede haber complejos niveles de cita; por ejemplo, puedo citar no sólo cómo le parece algo a John, sino cómo este parecer a John le parece a Mary. Pero no importa cuán compleja sea la cita, sigo siendo quien las hace; sigo siendo el centro citacional.

Cuando hablamos siempre contrastamos la forma en que nos parecen las cosas con la forma en que les parecen a otros. La forma en que las cosas parecen ser a los demás influencian la forma en que nos parecen a nosotros. Este suplemento de puntos de vista alternativos se niega cuando nos concentramos en inferencias deductivas en línea recta. La lógica de la deducción es una lógica para monólogos, una lógica ciclópea, de un solo ojo. Toda diversidad de puntos de vista queda filtrada. Sólo se admite lo que se sigue de nuestras premisas. E incluso en la lógica formal que trata de manipular casos no cubiertos por un conjunto específico de axiomas

(aun en las lógicas no monotónicas, que tratan de dar cuenta de situaciones y hechos que no se siguen de las premisas que se configuran en el sistema) seguimos limitados a inferencias que se ejecutan desde un solo punto de vista. Como ha escrito Raymond Reiter, «Todos los formalismos no monotónicos actuales tratan con agentes razonadores singulares. Sin embargo, está claro que con frecuencia los agentes deben adscribir inferencias no monotónicas a otros agentes, por ejemplo en el planeamiento cooperativo o en los actos de habla. Esos escenarios multiagentes requieren teorías formales apropiadas de las que actualmente carecemos».[15]

La restricción de la lógica a un solo punto de vista es una abstracción útil y legítima, pero se debe considerar limitada, pues no proporciona una imagen completa del pensamiento humano. En nuestro pensamiento natural, las opiniones de los otros ejercen influencia en las opiniones que sostenemos. No derivamos nuestras posturas sólo de los axiomas que aceptamos como verdaderos. Si la inteligencia artificial ha de emular al pensamiento natural, debe desarrollar programas que puedan manipular puntos de vista alternativos y no sólo razonamiento inferencial en línea recta. Debe desarrollar una lógica que de alguna manera tome en cuenta las expectativas y afirmaciones de un interlocutor y formalice un argumento conversacional, no sólo un argumento monológico. Esa expansión del pensamiento artificial ayudaría por cierto en la simulación de estrategias y situaciones competitivas. Por otro lado, si la cita se encuentra más allá de la inteligencia artificial, quizá sólo nosotros podamos ser los centros citacionales últimos en el pensamiento; quizá nuestras máquinas de pensar siempre serán citadas por nosotros, y nunca nos puedan citar en compensación.

Practicando distinciones

Otra clase de pensamiento distinto del pensamiento inferencial es la actividad de hacer distinciones.[16] Un programa de computadora puede hacer una distinción en el sentido de que puede seleccionar un ítem en lugar de otro, pero esa actividad presupone que los términos de la distinción ya han sido programados en la máquina. Una cuestión más elemental es si puede «nacer» una distinción en una máquina. ¿Puede una máquina establecer originalmente los términos de una distinción?

En nuestro pensamiento natural no inferimos distinciones. Reconocer que hay dos aspectos distintos en una situación es un acto de pensamiento más rudimentario que la inferencia. También es una señal de mayor inteligencia, especialmente si los dos términos de la distinción no se han establecido previamente en las nociones comunes almacenadas en nuestro lenguaje. Por ejemplo, apreciar que en una situación difícil hay algo amenazador y también algo insidiosamente deseable y tener cierto sentido del tono especial, tanto de la amenaza como de la atracción, es un acto crudo de intuición. No se deriva de premisas. Esta especie de pensamiento, este nacimiento de distinciones, está en el origen de las categorías que constituyen nuestro conocimiento común. Es anterior a los axiomas de los cuales se derivan nuestras inferencias.

En forma parecida, el almacén de reglas y representaciones que constituyen un programa de computadora, una base de datos y una base de conocimientos presupone que ya se han distinguido unas de otras las diversas representaciones almacenadas. Este almacén de distinciones ha debido ser construido por la inteligencia natural. Y cada representación, cada idea en la inteligencia natural, no ha sido sólo absorbida por la mente como un líquido es absorbido por un papel secante; cada idea ha debido distinguirse de las otras apropiadas.[17] Algún pensamiento y alguna distinción van en toda noción que tenemos. La instalación pensante de una idea siempre involucra distinción. ¿Hay alguna forma en que la inteligencia artificial pueda generar una distinción entre clases de cosas? ¿Pueden originarse distinciones en una máquina? ¿O la máquina pensante es como una mascota hogareña, que sólo se alimenta de lo que le damos?

Deseo

El deseo se halla involucrado con el pensamiento de dos maneras. Está primero el deseo de saber más: la curiosidad para aprender más hechos o el impulso a comprender con más plenitud. Pero también está el deseo hacia otras satisfacciones, como el alimento, el ejercicio, el reposo y cosas así. Llamaremos pasiones a estos deseos. ¿Cómo se relaciona el pensamiento con la pasión?

Una forma común de expresar esta relación es diciendo que la razón es la esclava de las pasiones.[18] En esta perspectiva, las

pasiones con las que nacemos establecen los fines que desearemos perseguir, las satisfacciones que buscamos; la razón entra en juego entonces para figurarnos cómo podemos obtener lo que las pasiones nos hacen buscar. Los deseos proporcionan los fines, el pensamiento proporciona los medios. En esta perspectiva queda poco espacio para la discusión racional de los fines porque los fines no son establecidos por la razón.

Esa comprensión de la relación entre el deseo y la razón encaja bien con algunas presuposiciones de la inteligencia artificial. Es fácil ver que la computadora puede ayudarnos a determinar cómo obtener un objetivo (quizás utilizando las técnicas del General Problem Solver iniciadas por Allen Newell, Cliff Shaw y Herbert A. Simon), pero la computadora tendría que tener los objetivos configurados de antemano, así como necesitaría tener configurados sus axiomas.[19] La computadora nos ayuda a obtener los fines elaborando las inferencias apropiadas al problema que enfrentamos y a los recursos que tenemos. De este modo, si la inteligencia natural es por cierto la esclava de las pasiones, la inteligencia artificial puede ir bien lejos en su reemplazo.

Pero la razón natural no es por completo externa a nuestros deseos. Es verdad que como agentes comenzamos con pasiones que preceden al pensamiento, pero antes que nuestro pensamiento entre en nuestros deseos y articule lo que buscamos, de modo que deseamos de una manera pensante. Deseamos no solamente alimentación, sino comer un almuerzo; no deseamos sólo refugio, sino un hogar. Nuestras pasiones quedan penetradas por la inteligencia. Más aún, surgen nuevas clases de deseos que sólo un ser pensante puede tener. Podemos desear honor, retribución, justicia, compasión, valor, seguridad contra futuros peligros, una sociedad política. Nuestro «deseo racional» involucra no sólo curiosidad y la articulación pensante de las pasiones sino también el establecimiento de formas de desear que no podrían ocurrir si no pensáramos.

La inteligencia artificial puede ser capaz de hacer algo con objetivos configurados de antemano, pero ¿puede emular la mezcla de deseo e inteligencia que constituye gran parte de lo que pensamos y hacemos? ¿Puede emular la curiosidad? La máquina pensante se mueve por energía eléctrica, pero ¿hay alguna forma de darle la clase de origen del movimiento que llamamos deseo? ¿Puede su razonamiento convertirse en un deseo pensante? ¿O todo deseo será para siempre el nuestro?

Derivar inferencias es una actividad intelectual que es menos radicalmente nuestra que las tres actividades que hemos examinado. Una vez que se han definido los axiomas y las reglas, cualquiera puede extraer conclusiones. Incluso si sucede que somos quienes desarrollamos las conclusiones, no necesitamos creer las conclusiones a las que llegamos. Sólo necesitamos decir que esas conclusiones se siguen de esas premisas. La inferencia puede permanecer en gran medida como algo sintáctico. Pero en la cita estamos más en lo nuestro, dado que distinguimos nuestro punto de vista del de alguien más. Al hacer una distinción también pensamos más auténticamente, con más independencia, pues vamos detrás de cualesquiera axiomas y premisas que alguien pueda definir en lugar nuestro y simplemente permitimos que una cosa se distinga de otra. En el deseo pensante expresamos el carácter que hemos desarrollado y la forma en que nuestras emociones han sido formadas por el pensamiento. La cita, la distinción y el deseo son formas de pensamiento más genuinas que la inferencia. Y aunque estas formas de pensamiento son más ampliamente las nuestras propias, no nos tornamos por ello meramente subjetivos o relativistas. Ellas expresan una objetividad y una verdad apropiadas a las dimensiones de pensamiento y ser en que se hallan involucradas, dimensiones subestimadas en el razonamiento inferencial.[20]

Si la inteligencia artificial fuera capaz de dar cuerpo a formas de pensamiento tales como la cita, la distinción y el deseo, parecería mucho más un reemplazo genuino de la inteligencia natural que un mero simulacro de ella. Parecería, en su artificialidad, ser similar a la luz artificial. Parecería que de alguna forma es capaz de originar su propio pensamiento, de hacer algo que no puede resolverse en el razonamiento de quienes hacen y usan las máquinas pensantes. Pero aun si la inteligencia artificial no puede dar cuerpo plenamente a esas actividades, puede al menos complementarlas, y precisamente al complementarlas nos puede ayudar a comprender lo que son. Podemos aprender muchísimo sobre la cita, la distinción y el deseo llegando a ver por qué no se los puede reproducir mecánicamente, si ese llega a ser el caso. Podemos aprender muchísimo sobre la inteligencia natural distinguiéndola de la artificial. Y si la inteligencia artificial nos ayuda a comprender lo que el pensamiento es (sea por emulación o por contraste) tendrá éxito en la empresa de convertirse no ya en una tecnología, sino en parte de la ciencia de la naturaleza.

Notas

[1] Alan Turing, «Computig Machinery and Intelligence,» *Mind 59* (1950):434-60.

[2] Frege habla de un *Gedankenbewegung* como del proceso que se supone que expresa su propia notación. Véase «On the Scientific Justification of a Conceptual Notation», en *Conceptual Notation and Related Articles*, editado por T. Bynum (Oxford: Claredon, 1972), 85.

[3] Véase Paul Harmon y David King, *Expert Systems in Business* (Nueva York: John Wiley & Sons, 1985).

[4] Jacques Arsac se pregunta «¿Cuántas actividades semánticas del hombre pueden representarse mediante signos en un lenguaje apropiado y tratadas 'informáticamente' [es decir, codificadas y manipuladas sintácticamente]? ¿Quién, a esta altura de los acontecimientos, puede determinar los límites que esta ciencia no será capaz de cruzar?» Arsac, *La Science Informatique* (París: Dunod, 1970), 45.

[5] Véase G.P. Baker y P.M.S. Hacker, *Frege: Logical Excavations* (Nueva York: Oxford University Press, 1984), 8: «El objetivo primario abogado por Frege fue sustanciar la tesis logicista de que la aritmética es parte de la lógica pura. Todo lo demás que hizo fue periférico. En consecuencia *él* vio lo que juzgamos su más grande logro, a saber, su invención de la escritura conceptual, como enteramente instrumental»

[6] La frase es de J.J. Scarisbrick, *The Reformation and the English People* (Nueva York: Blackwell, 1984), 74.

[7] Adapto estos ejemplos de Arsac, *La science informatique*, 34-47.

[8] Para una propuesta y compilación reciente del problema de la intencionalidad y la ciencia de la computación, véase Kenneth M. Sayre, «Intentionality and Information Processing: An Alternative Model for Cognitive Science,» *Behavioral and Brain Sciences* 9 (1986):121-65. Sobre la importancia de la «intencionalidad» o la «representación», véase la conclusión de Howard Gardner, *The Mind's New Science: A History of the Cognitive Revolution* (Nueva York: Basic Books, 1985), 381-92.

[9] En mi opinión, esta «sustancialización» de un «sentido» se presenta en algunas interpretaciones de la teoría del noema de Husserl: las que toman el noema como una representación mental que da cuenta del carácter intencional de la actividad mental. Ya he comentado esta cuestión y dado referencias sobre las diversas posturas y participantes en esta controversia en «Intentional Analysis and the Noema,» *Dialectica* 38 (1984):113-29, y en «Husserl and Frege,» *The Journal of Philosophy* 84 (1987)(de próxima aparición). Para un intento de explicar el significado sin recurrir a representaciones mentales, o, como se las llama a veces, «entidades abstractas,» véase mi ensayo «Exorcising Concepts,» *Review of Metaphysics* 40 (1987):451-63.

[10] Véase Ernest Fenollosa, *The Chinese Written Character as a Medium for Poetry*, ed. Ezra Pound (San Francisco: City Light Books, 1936), esp. p. 9: «Al leer chino no parecemos estar haciendo malabarismos con números, sino que parecemos contemplar cosas que hacen su propio destino».

[11] Como dice David Diringer a propósito del chino, «Hay una escasez extrema de estructura gramatical en el chino; hablando estrictamente, no hay una gramática china, y difícilmente haya una sintaxis», *The Alphabet*, 3a. ed. (Nueva York: Funk and Wagnalls, 1968), vol. 1, 63.

[12] Si tratamos de leer en voz alta algunas de las fórmulas desarrolladas por

C.A.R. Hoare nos encontraríamos hablando algo parecido al inglés pidgin. Véanse las fórmulas en *Communicating Sequential Processes* (Englewood Cliffs, N.J.: Prentice-Hall, 1985), 27-30, 43, 47-49. Los ideogramas tienden a expresar sucesos más que predicados, y el formalismo de Hoare es un intento de capturar sucesos en un proceso; véase pág. 25. Y Hoare está al tanto de que él se encuentra simbolizando no palabras o nombres, sino cosas y sucesos. El comienza su libro como sigue (pág. 23): «Olvidemos por el momento las computadoras y los lenguajes de computación, y en lugar de eso pensemos en los objetos del mundo que nos rodea, que actúan e interactúan con nosotros y entre ellos de acuerdo con algunos patrones de conducta característicos. Pensemos en relojes, en contadores, en teléfonos, en juegos y en máquinas expendedoras. Para describir estos patrones de conducta, primero decidimos qué tipos de sucesos o acciones serán de interés y elegimos un nombre diferente para cada clase.»

[13] Véase Baker y Hacker, *Frege: Logical Excavations*, pág. 35: La escritura conceptual de Frege «estaba designada para proporcionar una representación perspicua de las inferencias, para asegurar que ningún presupuesto quedara oculto. El corazón de la *Begriffsschrift es* entonces la elaboración de una notación para presentar inferencias y la definición de un sistema formal para verificar rigurosamente su poder de persuasión. El excluyó de su expresión en escritura conceptual todo 'lo que no tiene importancia para la cadena de inferencia'».

[14] Véase Robert Sokolowski, «Quotation», *Review of Metaphysics* 37 (1984):699-723.

[15] Raymond Reiter, «Nonmonotonic Reasoning», *Annual Reviews of Computer Science* 2 (1987): 183. Agradezco a John McCarthy por haberme llamado la atención sobre este artículo.

[16] Véase Robert Sokolowski, «Making Distinctions», *Review of Metaphysics* 32 (1979):639-76.

[17] Un ejemplo interesante de la forma en que un término puede basarse en varias distinciones, y del modo en que la «activación» de una u otra distinción puede modificar el sentido del uso concreto del término se encuentra en Pierre Jacob, «Remarks on the Language of Thought», en *The Mind and the Machine: Philosophical Aspects of Artificial Intelligence*, edición de S. Torrance (Nueva York: John Wiley, 1984), 74. «Para el uso que daba Bob al predicado [negro], algo contaría como negro si no era percibido como azul oscuro o cualquier otro color excepto negro, fuera o no teñido. Para el uso que daba Joe al predicado, algo contaría como negro no sólo si parecía negro sino también si resultaba no ser teñido». El sentido incidental de «no negro» hace una diferencia en el uso concreto de «negro».

[18] La frase es, por supuesto, de David Hume: «La razón es, y sólo debe ser la esclava de las pasiones, y nunca puede pretender otro oficio que el de servirlas y obedecerlas». A *Treatise of Human Nature*, edición de L. A. Selby-Bigge (Nueva York: Oxford University Press, 1960), vol. 3, 415.

[19] Para un resumen sobre el General Problem Solver (GPS) y el análisis de medios y fines, véase John Haugeland, *Artificial Intelligence* (Cambridge: MIT Press, 1985), 178-83.

[20] En *Artificial Intelligence*, Haugeland contrasta dos modelos de pensamiento: el «aristotélico», en el cual se dice que la mente piensa absorbiendo imágenes de las cosas, y el «hobbesiano», en el cual se dice que el pensamiento es la computación llevada a cabo en símbolos mentales. Haugeland llama a Hobbes «el abuelo de la IA» por su comprensión computacional de la razón (p. 23), pero llega a la conclusión de que podríamos tener que invocar una teoría del significado que involucre tanto

las imágenes como la computación (pág. 222). Me parece que en la filosofía de Husserl podemos encontrar ricos recursos para una teoría semejante, pues para él todas las presentaciones son articuladas y todas lás articulaciones mentales son presentacionales. Para Husserl, sintaxis y semántica son esencialmente partes de un todo más amplio. Contra Haugeland yo diría, sin embargo, que no se puede concebir que la mente absorbe imágenes de las cosas, sino que simplemente presenta las cosas en muchas formas diferentes.

4
Inteligencia artificial: un *aperçu*

Pamela McCorduck

Un paraíso imaginado: árboles de gruesos troncos con copas frondosas; viñas sinuosas; flores enormes. En ese paraíso, la pareja eterna: el brazo del hombre extendido, señalando hacia adelante mientras mira por encima de su hombro, quizá para empujar a su compañera, con el cabello al viento, mientras la abraza.

Otra escena: los árboles más altos, pesados por la lluvia reciente; ausencia notoria de brisa; las hierbas más densas ahora, pero aun así él trata de ocultarse (de ella, no de nosotros, porque alcanzamos a ver sus piernas por entremedio de la vegetación florida) mientras ella lo busca, dando sombra a sus ojos para ver mejor. Más escenas: ella se ha ido, luego él se va también; el artista ahora se concentra por completo en los detalles de la vegetación (figuras 1 a 3).

El artista nunca ha visto lo que sólo es imaginado; se ha unido a la larga tradición, desde las paredes de la cueva de Cro-Magnon hasta el arte funerario egipcio y Henri Rousseau, de expresar lo que puede ser, tan significativo para la imaginación humana como lo que es.

Excepto que el artista es un programa de computadora. Equipado (¿debo decir dotado?) con ideas sobre el crecimiento de las plantas, sobre el tamaño y forma de los seres humanos y las poses plausibles que pueden adoptar, equipado también con algunas ideas sobre arte (clausura, oclusión, equilibrio espacial, simetrías agradables o aburridas), el programa sigue adelante en forma autónoma, trazando dibujos por miles. Recuerda lo que ha hecho anteriormente, y no se repite a menos que se le pida hacerlo. Su nombre es AARON. Una pequeña broma sobre la vara de Aarón, según creí, pero mi creencia era errónea: se suponía que AARON

Pamela McCorduck. Conferencista en el programa de escritura en la Universidad de Columbia, autora de los libros Machines Who Think *(1979) y* The Universal machine *(1985).*



Figura 1

habría de ser el primero de una serie de programas, a ser nombrados alfabéticamente. En lugar de eso AARON ha perdurado, evolucionando en complejidad y madurez. Aaron es el nombre hebreo de Harold Cohen, el artista que creó el programa, quien lo dotó con su esencia y quien contempla, asombrado como cualquier otro, mientras AARON dibuja. AARON es inteligencia artificial.

Los dibujos de AARON suscitan cuestiones enigmáticas. Con toda seguridad, algunas de las mismas cuestiones que también suscitan otras obras de arte; cuestiones que tienen que ver con la naturaleza y significado del arte mismo, dentro de una cultura y fuera de ella; cuestiones sobre el papel del observador (veo un paraíso, pero otro puede ver una pesadilla vegetal, o, si no está al tanto de las convenciones del arte occidental, puede no ver nada en absoluto. Otros pueden incluso olvidar o disculpar la génesis de AARON, de modo de desaprobar lo que ven). La obra de AARON también une un conjunto de objetos de arte que reclaman por la identidad del artista (¿AARON? ¿Harold Cohen?).

AARON es sólo una máquina semiinteligente. Esto es, traza sus



Figura 2

dibujos pero no disputa con los críticos, con los dueños de galerías o incluso con Harold Cohen. No tiene aparato perceptual para «ver» lo que imagina (aunque ahora sabemos que lo que los seres humanos perciben es en gran parte una función de nuestras propias estructuras simbólicas internas, modeladas por un largo proceso de aculturación). Surge otra cuestión: careciendo de ojos, careciendo de intereses más allá de sus propios dibujos, ¿puede pretender AARON poseer en absoluto inteligencia? En otras palabras, la inteligencia ¿es todo o nada?

Para confundir más las cosas, he interpretado los dibujos a mi manera, pero ¿está la percepción sólo en los ojos del que ve? Supongamos que en lugar del paraíso lo que se dibuja es una Selva Negra tropical, y la pareja un Hansel y una Gretel tratando sin esperanza de volver de su abandono. ¿Es esto igualmente legítimo (o igualmente ilegítimo, dada la falta de interés de AARON en cualquier interpretación)?

Vamos ahora a esta cuestión. Por ahora, el punto es: AARON



Figura 3

es inteligencia artificial, inteligencia in vitro, no toda la conducta inteligente tal cual la reconocemos en seres humanos. En vez de eso, la inteligencia artificial es ciertas partes significantes de la conducta inteligente, cultivadas en silicio por las mismas razones que se cultivan y estudian las células: para comprender las partes como un paso hacia la comprensión del todo. Esta puede ser la instancia más problemática para quienes creen que la conducta inteligente, una vez separada de una inteligencia general plenamente funcional como la de la mente humana, no puede ser ya considerada en verdad inteligente.

Embrionarios como son, los hallazgos de la IA hacen más que servir para expandir nuestro conocimiento. Ya se han aplicado para asistirnos en una variedad de tareas, prácticas e inútiles. Quizá lo que es más importante, la IA ha comenzado a redefinir nuestro sentido de nosotros mismos y nuestro lugar en el mundo. Una disciplina sincrética, tan significativa para el arte como para la ciencia, para la emoción como para la razón, la IA es una parte imperiosa de nuestro pasado humano y una parte inevitable de

nuestro futuro humano (y, algunos dirán, extrahumano). No hay nada que se le parezca.

Durante los últimos treinta años la investigación en IA ha asumido dos estrategias generales, aunque ellas existen en reciprocidad. La primera estrategia ha consistido en imitar la inteligencia humana en un programa de computadora, en particular, encontrando modelos funcionales que eluciden el conocimiento humano. La segunda estrategia ha procurado atacar y resolver problemas sin necesidad de referirse a modelos de la inteligencia humana, con el propósito de exhibir comportamiento inteligente de un orden más alto, fuera del cráneo humano.

Como ciencia, la IA ha evolucionado en una búsqueda de los principios, quizá de las leyes de la conducta inteligente en general, sea los exhibidos por los humanos o por las computadoras. El primero de esos principios ya se ha propuesto; una tarea actual de la investigación en IA es verificarlo y articularlo más.

Otorgando la esencia

En los últimos años un objetivo fascinante ha capturado las imaginaciones a través de las disciplinas científicas. Uniéndose a los científicos computacionales, ingenieros, psicólogos y lingüistas que han perseguido este objetivo por casi treinta años, los médicos y biólogos han asumido de una forma científica el problema de cómo puede la mente surgir de la materia.

Durante casi tanto tiempo como el de los registros de que disponemos, los seres humanos han imaginado que imponían su esencia sobre los artefactos (ídolos, autómatas, robots, simulacros, deidades impredecibles, esclavos obedientes) inteligencias artificiales y animadas. Lo que es esta esencia humana, por supuesto, ha cambiado en el curso del tiempo, pero un tema es recurrente: ser humano es pensar, razonar, cogitar, asociar, crear.

Las primeras instancias de este impulso imaginativo son estructuras retóricas, y dada la centralidad del lenguaje, la forma persiste hasta hoy: historias, mitos, incluso argumentos filosóficos. Se las encuentra, por ejemplo, en la *Ilíada* de Homero: los sirvientes de Hefesto son «dorados, y en apariencia como jóvenes mujeres vivientes. / Hay inteligencia en sus corazones, y hay

lenguaje en ellas / y fuerza, y de los dioses inmortales han aprendido cómo hacer las cosas.»[1]

Relatos contemporáneos y muy parecidos aparecen en China: se dice que el aventurado intelectualmente (y físicamente) rey Mu de la dinastía Chou poseía un robot inmortal que estaba «muy cerca de la carne y la sangre artificial», todo parte de la gran fascinación china por la vida generada mediante procesos químicos. Esta idea encuentra nueva vida entre los árabes, nueve o diez siglos más tarde, en *Ilm al-takwin*, la «ciencia de la generación artificial», que con el tiempo condujo a la idea de *al-iksir*, o elixir de la vida, de los alquimistas medievales europeos.[2]

Esos árabes herederos de los helenos pueden haber sido los primeros en establecer formalmente que existe una diferencia entre las sustancias naturales y las artificiales. La distinción que hicieron los árabes no implicaba meramente que lo natural era superior a lo artificial, sino sólo que era diferente. Pero no muy diferente: ellos afirmaban que un medio excelente para conocer lo natural era estudiar lo artificial. Aunque esta premisa subyace a toda la idea occidental de modelización científica, no habría tenido mucho sentido para los chinos tradicionales, que se inclinaban a ver todo —animado e inanimado, natural y artificial— como conectado indisolublemente. Esta visión del mundo, tan fluida e inclusiva, penetraba tanto al sintoísmo como al budismo, y luego aisló a los japoneses modernos de los debates sobre si la inteligencia artificial es real o falsa. En ese tema ellos nos contemplan asombrados.

Pero volvamos a la historia occidental. La Europa medieval parece obsesionada por las inteligencias artificiales: ellas aparecen en todas partes donde se ejerce el intelecto. El papa Silvestre II, Alberto Magno, Roger Bacon y otros fueron de la compañía de la cual se dijo que habían hecho cabezas de bronce que decían el futuro y resolvían complejos enigmas, cabezas broncíneas que eran tanto prueba como fuente de sabiduría. El místico español Ramón Lull compró en oferta un artefacto árabe, una máquina pensante llamada *zairja*, y lo reconstruyó sobre líneas más cristianas, llamándolo, sin indebida modestia, *Ars Magna*. Consistía en una serie de discos concéntricos que podían equipararse a categorías u otros criterios, siendo la idea dar razón de cualquier tema. El *Ars Magna* y su antecesor árabe, la *zairja*, estaban modelados sobre el supuesto de que el pensamiento podía desarrollarse fuera de la mente humana y, por tanto, podía mecanizarse.

La mente de la estrategia filosófica

Si la mente es una propiedad (o proceso) humana esencial, su naturaleza es elusiva. En el siglo diecisiete se desarrollaron violentas disputas sobre la mente y el cuerpo, sobre si son diferentes o son lo mismo. Descartes concluyó que eran diferentes; Spinoza rechazó este dualismo, creyendo que la mente y el cuerpo humano eran dos aspectos de la misma cosa, dos atributos de Dios. En 1650, el año en que murió Descartes, el arzobispo Ussher publicó su famoso cálculo del comienzo del mundo: lo ubicó en el 4004 a.C. Protociencia y seudociencia coexistían dificultosamente. Leibniz también ponderó la cuestión mente-cuerpo y con el tiempo decidió que la mente y el cuerpo eran por cierto mónadas separadas, pero exactamente coincidentes. Bajo la influencia de los elegantes mecanismos de Newton, filósofos como Locke y Hume se aplicaron a la búsqueda de una comprensión racional de la mente, tratando de adivinar leyes análogas a las de Newton, pero no de la materia sino del pensamiento.

Esto tenía sus riesgos. La Iglesia, por ejemplo, con sus fuertes ideas sobre una adecuada actitud cristiana y dualista acerca de la mente y el cuerpo, no daba la bienvenida al disenso, como Julien Offray de la Mettrie descubrió un siglo después de Descartes. La Mettrie estudió a los filósofos y los escarneció: palabras sin sustancia, se burlaba, especulación sin conocimiento. En breve, pura retórica.

En 1747 La Mettrie publicó un libro llamado *L'Homme Machine*, basado en su práctica como médico. Proponía una teoría comprehensiva de la mente y citaba evidencia de que ciertas sustancias físicas afectaban al pensamiento; dieta, preñez, drogas, fatiga, enfermedad, todo figuraba en sus análisis. Rudo y peleador, pretendía que sus teorías (y su lenguaje) sacudieran. Se hizo de enemigos, no sólo entre las autoridades, que lo echaron primero de París y luego de los Países Bajos, sino también entre los filósofos que finalmente lo recibieron en la corte de Federico el Grande, en Berlín. Después de su muerte, ellos desecharon sus ideas y rehusaron incluso mencionar su nombre.

La Mettrie es importante porque es el primero en ofrecer evidencia empírica de la teoría: las revoluciones de las ciencias físicas lo tocan en forma diferente de la forma en que él tocó a los

filósofos. Su obra marca el comienzo del fin del amateurismo en la comprensión de la mente humana.

La computadora

Recordar los intentos tempranos de desarrollar inteligencias artificiales (capturar la mente fuera del cráneo humano) no equivale a decir que no hay nada nuevo bajo el sol. La computadora, el primer instrumento capaz de procesar símbolos de propósitos generales, comienza torpemente (luego sigue con un poco más de fluidez) a capturar ciertas cualidades esenciales del pensamiento humano.

Patrick Winston, del MIT, ha observado que la computadora es un sujeto experimental ideal, requiere poco cuidado y alimentación, es infinitamente paciente y no muerde.[3] La computadora permite definir, construir y testear el comportamiento inteligente de una manera rigurosa y proporciona la rápida retroalimentación necesaria para progresar en la experimentación. De este modo, por ejemplo, la noción de *símbolo* toma un significado preciso y demuestra ser esencial para el pensamiento inteligente.

La computadora ha hecho explícita la división fundamental entre los dos componentes de la conducta inteligente, el *hardware* y el *software*, desmitificando por fin el acertijo mente-cuerpo. La computadora no sólo proporciona una instancia de funcionamiento simbólico que surge de la materia; también revela cómo puede ser que esto suceda.

En 1948 el matemático del MIT Norbert Wiener publicó un trabajo breve pero seminal titulado *Cybernetics*. Este trabajo registró el paso de un paradigma dominante, la energía, a otro nuevo, la información. Una importante ventaja del nuevo paradigma para explicar el pensamiento era que trataba con sistemas abiertos, sistemas acoplados al mundo exterior tanto para la recepción de impresiones como para la ejecución de conductas; el viejo paradigma de la energía sólo trataba con sistemas cerrados y conservativos. Otra ventaja del nuevo paradigma, quizá más importante, era que trataba de la conducta de símbolos, que pronto emergerían como algo central para el estudio de la acción inteligente. El pequeño libro de Wiener apenas hace alguna mención de la computadora en su primera edición (un lapsus poco

sorprendente, dada la torpeza y poca confiabilidad de las máquinas de aquel entonces).

Que la computadora era potencialmente un manipulador de símbolos de propósitos generales y que podía usarse como tal era precisamente la opinión compartida por todos los investigadores tempranos de la IA. Durante el verano de 1956 diez de ellos se reunieron, durante períodos más o menos largos, en el campus del Dartmouth College. Como dijeron a su benigno patrón de la fundación, trabajaron sobre la base de la conjetura de que todo aspecto del aprendizaje o todo rasgo de la inteligencia se puede en principio describir tan precisamente como para hacer que una máquina lo simule.[4]

En los años de la conferencia de Dartmouth existía ya un programa que hizo época. Obra de un equipo de la Carnegie Mellon University (entonces Carnegie Tech) y de la Rand Corporation (incluyendo a Allen Newell, J.C. Shaw y Herbert Simon), el programa había sido llevado a la conferencia por Newell y Simon. Llamado el Logic Theorist, probaba ciertos teoremas de *Principia Mathematica* de Whitehead y Russell. El Logic Theorist descubrió una prueba para el teorema 2.85 más breve y más satisfactoria que la usada por Whitehead y Russell. Simon escribió las buenas nuevas a Lord Russell, quien respondió encantado. Pero *The Journal of Symbolic Logic* declinó publicar el artículo (cuyo coautor era el Logic Theorist) que describía la prueba.

El Logic Theorist, creado en un medio intelectual muy diferente del que compartían otros en Dartmouth ese verano, fue inventado por personas no demasiado comprometidas con la lógica, los formalismos matemáticos, las redes neuronales o cualquiera de los intentos tempranos de hacer que las máquinas pensaran. En lugar de eso, el Logic Theorist descansaba sobre la así llamada perspectiva del procesamiento de la información, que sostenía que los sistemas complejos de procesamiento de la información, construidos sobre componentes relativamente escasos y sencillos, podían exhibir comportamiento inteligente. Hacer que esos procesos trabajaran requería un conocimiento íntimo de las computadoras; Newell, Simon y sus primeros estudiantes de IA eran tan hábiles como inventivos con lenguajes de computadora como lo eran en electrónica. Diseñaron un lenguaje de programación de alto nivel llamado IPL-V (Information Processing Language Five) que reflejaba lo que la psicolo-

gía cognitiva ya había demostrado sobre la memoria asociativa humana.

John McCarthy, uno de los organizadores originales de la conferencia de Dartmouth e inventor de la frase *inteligencia artificial*, gustó de la idea general del procesamiento de listas, pero estaba ofendido por la desprolijidad de IPL-V. McCarthy creó su propio lenguaje, LISP (por List Processing). En sus muchos dialectos, LISP devino la lingua franca incomparable de la investigación y de las aplicaciones en IA en los siguientes veinticinco años. No sólo los investigadores en IA comenzaron a producir programas que podían ejecutar ciertas tareas que se consideraba requerían inteligencia, sino que comenzaron a articular un conjunto de ideas sobre la inteligencia que eran más revolucionarias por sus implicaciones que cualquier programa específico que jugara al ajedrez o demostrara teoremas.

En 1975, en ocasión de recibir el Premio Turing (el premio más prestigioso en ciencias de la computación), Allen Newell y Herbert Simon articularon, bajo la forma de una hipótesis científica, un supuesto que ellos creían era subyacente a todo trabajo en IA. Hablaron de una hipótesis de un sistema físico de símbolos.[5]

Toda ciencia —explicaban— caracteriza la naturaleza esencial de los sistemas que estudia. Estas caracterizaciones son cualitativas (la doctrina de la célula en biología, la tectónica de placas en geología, el atomismo en la química) y establecen un marco dentro del cual se pueden desarrollar estudios más detallados, a menudo cuantitativos. En computación, la descripción cualitativa que forma y define la ciencia es un sistema físico de símbolos.

Lo de *físico* denota claramente que esos sistemas obedecen a las leyes de la física: son realizables mediante sistemas de ingeniería hechos con componentes de ingeniería. Un *sistema de símbolos* es una colección de patrones y procesos, siendo los procesos capaces de producir, destruir y modificar los símbolos. La propiedad más importante de los patrones es que son capaces de designar objetos, procesos u otros patrones. Cuando los patrones designan procesos, pueden interpretarse. La interpretación involucra llevar adelante los procesos diseñados.

Esta perspectiva proporciona el marco conceptual para estudiar la inteligencia (humana o no), que puede definirse más precisamente como la capacidad de procesar símbolos. De este modo, los seres humanos y las computadoras son dos miembros

de una clase más amplia definida como procesadores de información, una clase que incluye muchos otros sistemas procesadores de información —económicos, políticos, planetarios— y, en su generalidad, una clase que amenaza abarcar el universo. Los seres humanos y las computadoras, sin embargo, son no sólo procesadores de información sino también agentes inteligentes: son capaces de computar algo que está más allá de sus propios procesos.

¿Cómo llegaron Newell, Simon y sus colegas a estas definiciones de forma y función de la inteligencia? Estaban en posición para hacerlo porque tenían para estudiar una máquina física llamada computadora; podían observar (y alterar) la conducta de la máquina, siguiendo el rastro de las conexiones explícitas de un nivel a otro a través de su jerarquía arquitectónica, desde los materiales semiconductores y metálicos hasta el nivel global de configuración, cada nivel explícitamente conectado a los de arriba y a los de abajo.

Los supuestos de trabajo del sistema físico de símbolos han llevado a dos hipótesis importantes pero ligeramente distintas sobre la naturaleza de la mente. Una de ellas, que postula un agente inteligente trabajando en un medio llamado nivel de conocimiento, es de Newell, y surge directamente de la hipótesis del sistema físico de símbolos. Generalmente es de diseño jerárquico. La otra, de Marvin Minsky, del MIT, utiliza los mismos supuestos básicos del sistema físico de símbolos, pero postula una «sociedad» de agentes (algunos inteligentes, otros no) que trabajan más heterárquicamente para producir los resultados que llamamos mente.

La mente, sugiere Allen Newell, es un agente inteligente, un miembro de la clase especial de procesadores de información que computan algo más que sus propios procesos. Una inteligencia como la que la mente exhibe es resultado de una conducta aditiva de una jerarquía de funciones, comenzando por el nivel más primitivo y elaborando hasta el nivel más complejo.[6]

Newell puede identificar correspondencias específicas entre los niveles sistémicos de un ser humano como un agente inteligente y los de un sistema de computadora, cada nivel sistémico con su propio medio físico y sus reglas de operación. El más básico es el nivel de dispositivo, hecho de materiales semiconductores en las computadoras y materiales bioquímicos en los humanos; luego

viene el nivel de circuitería, hecho de resistencias, inductores y capacitores en las computadoras y de células en los humanos. Se ha pensado generalmente que el nivel más alto es el de configuración: en las computadoras, la conducta global de memorias, cintas, discos y dispositivos de entrada-salida; en los humanos, el cuerpo que contiene el cerebro.

Pero quizás, especula Newell, debe agregarse un nivel por encima del nivel de la configuración, más allá del cuerpo individual que contiene el cerebro. Newell lo llama el nivel del conocimiento, y siguiendo su anterior taxonomía, el sistema a ese nivel es el agente; los componentes son objetivos, acciones y cuerpos. El medio a nivel del conocimiento es, por supuesto, el conocimiento.

De este modo un agente —un humano o un sistema de computadora— posee conocimiento (codificado, sin embargo, al nivel de los símbolos, no al del conocimiento) que procesa para determinar las acciones a seguir. El conocimiento requiere tanto estructuras como procesos: una forma de conocimiento, estructural, es impotente sin la otra clase, procedimental. La conducta a nivel del conocimiento se halla regulada por el principio de racionalidad; se seleccionan acciones para alcanzar los objetivos del agente. La racionalidad, sin embargo, no es perfecta sino limitada.

En un libro reciente dirigido a los no especialistas y titulado *The Society of Mind*, Marvin Minsky ofrece otra perspectiva de lo que puede ser la mente, ideas destiladas de toda una vida profesional dedicada a la investigación en IA.[7] Mientras que comparte el supuesto de la IA de que la inteligencia surge de la materia (o no-inteligencia), él adopta sobre todo ese proceso una perspectiva menos jerárquica que la de Newell.

Minsky sugiere que la entidad llamada mente es el producto de muchos agentes trabajando a veces juntos, a veces en conflicto, cada uno con objetivos mayores o menores y conocimientos superficiales, pero produciendo en conjunto todos los fenómenos de los que acostumbramos a hablar con respecto a la mente: el sujeto, la individualidad, la comprensión y la introspección, memoria, inteligencia, aprendizaje, etcétera. En un delicioso y poderoso metaejemplo, el libro mismo se compone de muchos ensayos breves, a menudo de menos de una página, formando en total una instancia del punto más importante. (En un epílogo, Minsky dice que trató de escribirlo en forma diferente, pero que no

funcionó. «Una mente es demasiado compleja para encajar en el molde de las narrativas que comienzan *aquí* y terminan *allá*; un intelecto humano depende de las conexiones en una compleja madeja, que simplemente no funcionaría si estuviera netamente desenredada.»[8])

En esos breves ensayos Minsky trata con casi todos los aspectos del pensamiento, que «no está basado en una forma simple y uniforme de 'lógica', sino en miríadas de diversas clases de procesos, argumentos, estereotipos, críticos y censores, analogías y metáforas. Algunos de ellos se adquieren mediante la operación de nuestros genes, otros se aprenden a partir del entorno, y otros más son construidos por nosotros mismos. Pero aun allí dentro de nuestras mentes, nadie en realidad aprende solo, pues cada paso emplea tanto cosas que hemos aprendido antes del lenguaje, la familia, los amigos, como de nuestros propios estados del ser anteriores. Sin que cada etapa enseñe a las siguientes, ninguna persona podría construir algo tan complejo como una mente».[9]

¿Piensa la gente lógicamente? No en realidad. ¿Lo hacen las computadoras? Tampoco. En lugar de eso, ambas emplean conexiones, procesos que involucran causas, similitudes y dependencias. Las conexiones entre los agentes son miríadas, a menudo indirectas. Minsky duda de que las leyes científicas de la inteligencia que surjan eventualmente puedan ser tan simples como las leyes de la física, porque la inteligencia (manipulación de símbolos) es más complicada que la materia.

La ciencia de la inteligencia artificial

Desde los comienzos de la computación moderna se ha creído ampliamente que si una persona ordinaria se pudiera comunicar con las computadoras en lenguaje ordinario, natural, se facilitarían muchos problemas severos de la comunicación. (Por cierto, esta creencia continúa: una porción significativa del esfuerzo en el proyecto japonés de la quinta generación para desarrollar computadoras inteligentes en gran escala se concentra en el procesamiento del lenguaje natural.[10]) Las estrategias computacionales iniciales frente al problema de la comprensión del lenguaje natural residían en modelos del lenguaje formales, virtualmente

matemáticos, y cualquiera haya sido su elegancia, los resultados que produjeron han sido insatisfactorios. El problema de la traducción automática entre lenguajes naturales, por ejemplo, uno de los primeros problemas no matemáticos abordados en la computación de posguerra, creó grandes esperanzas y luego las defraudó; ha habido altos y bajos episódicos desde entonces.

Los investigadores del lenguaje en IA han asumido cierto número de estrategias, y una de las más influyentes ha sido la de Roger Schank y sus colegas en Yale. A principios de los años 70 Schank formuló una teoría de la *dependencia conceptual*, que sugiere que los conceptos, no las palabras, son los primitivos apropiados con los que los lingüistas deben tratar (conceptos tales como posesión-cambio-acción en lugar de *dar, tomar, comprar* o *vender*). Schank propone un sistema de representación universal a nivel de la dependencia conceptual. A ese nivel, cuando dos frases son idénticas en significado, a despecho del lenguaje, hay sólo una representación. Este trabajo, en su atención a los contextos más amplios, tiene poco interés en concentrarse en simples interpretaciones frase por frase.[11] Schank reporta que cada implementación de su teoría general ha abierto problemas nuevos, no anticipados, y ha refinado y expandido las premisas en que la teoría se basa.

Sin embargo, la investigación del lenguaje en IA (y los schankianos sólo representan una estrategia) ha recorrido un largo camino. Hace treinta años los estudiosos se hubieran burlado: ¿cómo puede comprender una computadora la diferencia entre *la pluma está en la caja* y *la caja está en la pluma*? La respuesta es que un programa inteligente puede a menudo resolver la ambigüedad lingüística, parafrasear una narración simple, inferir respuestas a preguntas sobre ella e incluso narrar historias sencillas.

A medida que las computadoras proliferaron después de la Segunda Guerra Mundial y su potencial de procesamiento de la información impresionaba a un pequeño número de investigadores, algunos encontraron irresistible el paralelismo entre la naturaleza activado-desactivado de la neurona y el conmutador electrónico. A comienzos de la década de 1940 un brillante neurofisiólogo de la Universidad de Illinois, Warren McCulloch, se unió a un joven matemático, Walter Pitts, para intentar definir la mente matemáticamente como una red de neuronas interconectadas, dispositivos de tipo activado-desactivado. El trabajo de McCulloch-Pitts, pese a que

fue influyente durante una década o más, perdió el favor de la comunidad de IA. Muchos se alejaron de él, prefiriendo una perspectiva más cercana al procesamiento de la información.

Treinta años más tarde, con la disponibilidad de nuevas computadoras poderosas, la idea de las redes neuronales surgió de nuevo, vitalizada dramáticamente y llamada conexionismo. Pese a que el conexionismo o investigación en redes neuronales pretende construir un modelo de la inteligencia natural utilizando componentes de computadora, pocos son ahora los que comparan las neuronas con los conmutadores de una computadora.

Vale la pena repetir que todas las estrategias en inteligencia artificial comparten tres supuestos: que la conducta inteligente es explicable en términos científicos, que puede tener lugar fuera del cráneo humano y que la computadora es el mejor instrumento de laboratorio para estudiar esas proposiciones. Qué estrategia en particular probará ser la mejor está por ser visto.

A mediados de la década de 1960, un pequeño grupo de investigadores de Stanford, conducidos por Edward Feigenbaum y Joshua Lederberg (hoy presidente de la Rockefeller University), decidieron transformar su impaciencia hacia los problemas de juguete que habían acaparado los esfuerzos de los investigadores en IA —ajedrez, juegos abstractos, y cosas así— en una exploración de problemas del mundo real. Su objetivo era tratar de simular la inducción científica: ¿cómo es que los científicos, confrontados con un problema, razonan su camino hacia una solución? Como prueba seleccionaron el difícil problema del análisis espectrográfico, un procedimiento complicado para analizar la composición de moléculas orgánicas, entonces dominio de químicos con grado de Ph. D.

DENDRAL, el programa que a su tiempo salió de este esfuerzo, fue el primer sistema experto, un método para capturar la experiencia humana que se ha probado en diversos dominios, ocasionando considerable excitación en el mercado a medida que las empresas comerciales descubrían las ventajas de las aplicaciones de sistemas expertos.

En principio, el sistema experto es simple. Consiste en tres subsistemas interconectados:

1. Una *base de conocimientos*, que comprende hechos, supuestos, creencias, heurísticas («conocimiento experto») y métodos

para tratar con la base de datos para alcanzar los resultados deseados, tales como un diagnóstico, una interpretación o una solución a un problema.

2. Una *base de datos*, una colección de datos sobre objetos y sucesos sobre la cual la base de conocimientos ha de operar para alcanzar los resultados deseados.

3. Una *máquina de inferencia*, que permite trazar inferencias a partir de las interacciones entre el conocimiento y las bases de datos.

Se presenta un problema al sistema, que solicita datos de quien efectúa la consulta y, eventualmente, ofrece consejos para llegar a la solución. A veces se puede preguntar al sistema en lenguaje natural, lo que hace que sea no solamente valioso para profesionales ocupados, sino que lo torna un perfecto dispositivo de instrucción para los novicios. De esos comienzos relativamente simples ha surgido un negocio internacional de muchos millones de dólares que promete pronto devengar muchos más.

La inteligencia artificial, al afrontar como lo hace intereses importantes de los seres humanos, sus capacidades simbólicas, no puede menos que afectar a muchos otros dominios. ¿Quién puede decir cuál habrá de ser el más importante? La elucidación de la inteligencia humana es importante sin duda; lo mismo, sin embargo, es su incremento o su ampliación, con la ayuda de asistentes mecánicos inteligentes. Pero hay otros terrenos, menos obvios, que pronto podrán manifestar su influencia. Consideremos, por ejemplo, la actual situación en las artes visuales. Muchas de las bellas artes se han alejado hace tiempo del acceso y la comprensión cotidiana. Los críticos nos aseguran que si tuviéramos discernimiento y práctica, podríamos ver por qué un trozo de hierro del estudio de Richard Serra es diferente de un trozo de hierro listo para ser arrastrado a la empresa de metales de Joe Scrap, por qué las pinturas de Andy Warhol sobre las latas de sopa Campbell o las cajas de Brillo son más significativas que las latas y las cajas mismas, por qué es digno de nuestra atención que el difunto Joseph Beuys depositara un bollo de manteca en el piso del Museo Guggenheim y lo llamara arte.

Los críticos están en lo cierto. Sin embargo, a veces sospechamos

un gigantesco juego de embaucamiento, y nuestras sospechas se alimentan de las disputas monumentales entre los mismos expertos acreditados. Parece que el discernimiento y el entrenamiento, por sí solos, no son suficientes. Si los de afuera están confundidos, el discurso profesional sobre el arte (en tan gran medida histórico-descriptivo, tan a menudo frustrantemente inexacto) parece orillar lo arbitrario, y es fácil sospechar abusos de confianza.

Vuelvo ahora a AARON y a su obra, una conclusión apta para un ensayo sobre la inteligencia artificial. Cuando Harold Cohen comenzó a trabajar con AARON, deseaba «comprender mejor la naturaleza de los procesos de creación artística que lo que permite la misma práctica del arte, pues bajo circunstancias normales el artista proporciona un ejemplo casi perfecto de un cuerpo de conocimiento obviamente presente, pero virtualmente inaccesible.»[12] Los dibujos de AARON no pretendían específicamente ser estéticamente placenteros, aunque el programa era capaz de generar dibujos que lo fueran. El objetivo era permitir examinar ciertas propiedades del dibujo a mano alzada, lo que Cohen llama, en una frase deliberadamente general, «estar-en-lugar-de-otra-cosa» *[standing-for-ness]*.

Específicamente, Cohen buscaba una respuesta más precisa a ciertas cuestiones importantes ¿Qué es una imagen? ¿Cómo pueden marcas bidimensionales en una superficie evocar en la mente humana objetos reales del mundo? ¿Existen ciertos universales en nuestras estructuras cognitivas que permiten que los humanos infieran significados de imágenes bidimensionales? En síntesis, Cohen buscaba no sólo la gramática de las imágenes bidimensionales (figuras cerradas en lugar de objetos sólidos, el cerramiento en lugar de la tercera dimensión de las relaciones espaciales) sino también los comienzos de su semántica, las condiciones mínimas para que un conjunto de marcas funcionara como una imagen. Las primeras imágenes de AARON eran abstractas, compartiendo con el arte paleolítico (y quizá con la mayor parte del diseño humano de imágenes) el intento de minimizar el problema de la dimensión. Nada hay implícito sobre la profundidad en los espacios entre elementos pictóricos. AARON se movió luego al estudio de los bordes y de la profundidad implícita.

Recientemente el programa ha sorprendido a sus veteranos observadores haciendo dibujos figurativos, los que Harold Cohen el pintor evitaba, pero a los que Harold Cohen el estudioso del

diseño de imágenes y arte sentía necesario explorar. «La primera vez que el programa acumuló formas cerradas en algo que parecía ser una aproximación a una figura, y encontré un montón de casi-gente contemplándome sin ojos desde mi vieja 4014, temblé de miedo. ¿En qué me había metido?»[13] Esa casi-gente es distintiva de AARON: retozando en jardines tropicales fantasmagóricos pero reconocibles, en campos de juego en la playa o (quizás en homenaje a Cézanne) en mallas de baño, se la puede identificar por su género.

Para repetirlo, AARON no puede «ver» figuras humanas, árboles, hojas, campos de juego ni cualquier otra parte del mundo material. En vez de eso, Cohen le ha proporcionado a su programa ideas sobre esas cosas, conocimiento indirecto. A la manera de un artista que es requerido para que conceptualice lo que un viajero le dice sobre un lugar lejano que nunca ha visitado, o a la manera de un artista que ilustra un cuento de hadas, AARON se ha dejado llevar. Las figuras humanas aparecen en muchas poses; AARON sabe cómo se mueven las junturas, cómo las figuras humanas conservan su equilibrio. Es capaz de evocar un repertorio casi infinito de poses humanas plausibles. Cuando esas figuras se colocan en un paisaje, las ideas generales que tiene AARON sobre el crecimiento de las plantas le permiten generar una gran variedad de plantas y árboles individuales.

En AARON se ejemplifica una idea central de la inteligencia artificial: el programa es capaz de generar la ilusión de un conjunto completo y coherente de imágenes a partir de una representación de orden más bajo, comparativamente simple y dispersa. Pero si AARON puede funcionar como una representación de la actual investigación en IA, también representa, desde mi punto de vista, un ejemplo de la sorprendente influencia que la IA puede tener en campos muy distantes de sus orígenes en las ciencias matemáticas y la ingeniería.

Lo que AARON hace es claramente diseño de imágenes, pero ¿es arte? Sí, nos contesta Cohen serenamente, apartándose de su creación, su protegido, como lo llama, que se entretiene con sus dibujos. «En la cultura occidental —prosigue— siempre atribuimos el más alto nivel de responsabilidad —y elogio o culpa— al individuo que trabaja en el más alto nivel conceptual. Podemos escuchar cien ejecuciones diferentes de un cuarteto de Beethoven sin dudar siquiera que hemos escuchado a Beethoven. Recordamos los nombres de los arquitectos, no de los albañiles que

hicieron sus edificios. Y particularmente valoramos a aquellos cuyo trabajo deja al arte en un estado distinto del estado en que lo encontraron.»[14]

El programa AARON, según él cree, está en relación con sus dibujos individuales en la misma forma en que los ideales platónicos están en relación con sus instancias mundanas. Es un paradigma. Que Cohen haya encontrado una forma de hacer su voluntad a través del paradigma antes que mediante una instanciación singular significa simplemente que su nivel de compromiso es mucho más elevado, conceptualmente hablando, de lo que ha sido posible antes para un artista visual. Esto se parece a la forma en que un compositor escribe una partitura, en lugar de brindar una ejecución, aunque en AARON el programa es responsable por todas las ejecuciones. Es como si una partitura pudiera ejecutarse a sí misma (un deseo de los compositores a lo largo de este siglo).[15]

El arte, alega Cohen, es el ejercicio más variado y sutil de la humanidad en la representación del conocimiento. La historia del arte no es simplemente el registro de los cambios en el significado o el estilo, sino de los cambios en la relación entre significado y estilo, todos ellos harto más pequeños dentro de una cultura específica que entre las culturas.

¿Un código simbólico universal?

Tenemos la capacidad de comprender épocas que no son las nuestras, culturas que no son la propia. Llevemos esto mucho más allá, al reino de la especulación y, tomando prestado el continuum de la comprensión lingüística de Schank, digamos que hay épocas en las que encontramos un artefacto de una cultura o de un tiempo que no son los nuestros que posee sentido, que nos permite alcanzar una comprensión cognitiva más profunda, que trabaja en pro de una comprensión empática más honda.

Pero con alguno de los modernismos y con lo que le ha seguido, a menudo quedamos confundidos; no necesitamos depender de expertos que no se ponen de acuerdo entre ellos. Carecemos de un lenguaje objetivo, preciso y coherente para definir, medir o aunque sea interpretar los procesos y productos de la representación, en este caso imágenes visuales.

Cohen, para citar a alguien, duda acerca de la existencia de ese código, desechando el concepto como «el modelo telecomunicacional del arte.»[16] El sugiere que la transacción entre el hacedor de imágenes y el lector de imágenes tiene lugar en un nivel simple de conocimiento; el sentido de significatividad se genera a través de la estructura de la imagen, más que por su contenido. De este modo, la interpretación de los tres dibujos ofrecidos a comienzos de este ensayo se basa en la vida de aculturación de un individuo, y en nada más. Cohen me permitiría mis interpretaciones aculturadas, satisfecho de que la fuerza pictórica generativa de AARON me haya llevado a una transacción que no es una imagen de dos personas en un paraíso imaginado, sino más bien el registro pictórico de un acto de voluntad de AARON y, finalmente, de Harold Cohen. Y esto está de acuerdo con la creencia de cierto número de prominentes estetas y filósofos de que esto es todo lo que se puede decir con confianza sobre cualquier obra de arte.[17]

Cohen bien puede tener razón para dudar de la universalidad de las formas representacionales más elevadas (al fin y al cabo Freud y Jung no tuvieron éxito en esa búsqueda). Al mismo tiempo él muestra universalidad a un nivel cognitivo más bajo, donde ciertos motivos son ubicuos en las expresiones humanas (zigzags, cruces, cuadrados, mandalas, combas, etc.) y están todos construidos sobre la base de elementos aun más simples.

Pero supongamos que no tiene razón. Supongamos que a algún nivel, todavía por definirse, un conjunto de conceptos universales subyace a todas las expresiones simbólicas, siendo las artes visuales sólo un aspecto de esto. ¿Se puede elucidar ese nivel? Si es así ¿clarificará el impulso humano a expresar las cosas simbólicamente? ¿Sugerirá esto que no podemos realmente hablarnos significativamente entre nosotros o pretender que lo hacemos? La expresión precisa, después de todo, ha introducido una cierta especie de mutismo transdisciplinario en la ciencia, aun cuando también ha introducido un conjunto de universales más amplio.

La inteligencia artificial se ha propuesto muchas metas ambiciosas: una comprensión rigurosa de la inteligencia donde quiera que se manifieste, incluyendo evidencia más plena para la hipótesis del sistema físico de símbolos; conceptos más precisos de la mente, la comprensión, el aprendizaje, la representación del conocimiento y los usos del lenguaje natural. Si las dudas de Harold Cohen son infundadas (si la inteligencia artificial puede

comenzar a echar luz sobre un código universal, si tal cosa existe, un nivel de significado subyacente a las más importantes expresiones simbólicas de la experiencia humana), entonces las preguntas ya formuladas en estudios de elocuciones lingüísticas y representaciones visuales serán patentes también en otros lugares. Si una pregunta semejante ha de moverse más allá de la retórica (una frase para desalentar a cualquier escritor) y ha de ser contestada con conocimiento empírico, la inteligencia artificial —entre todas las cosas— es nuestra mejor esperanza.

Notas

Figuras 1-3 reproducidas por cortesía de Harold Cohen.

[1] Excepto cuando se lo señala, estos y otros materiales sobre la historia temprana de la inteligencia artificial, con sus referencias, se presentan en Pamela McCorduck, *Machines Who Think* (San Francisco, W.H. Freeman & Co., 1979).

[2] Joseph Needham, *Science in Traditional China* (Cambridge: Harvard University Press, y Hong Kong: Chinese University Press, 1981).

[3] Patrick Henry Winston, *Artificial Intelligence* (Reading, Mass.: Addison-Wesley, 1977).

[4] McCorduck, *Machines Who Think*.

[5] Allen Newell y Herbert Simon, «Computer Science as Empirical Inquiry: Symbols and Search», *Communications of the Association for Computing Machinery* (marzo de 1976).

[6] Allen Newell, «The Knowledge Level», *Artificial Intelligence* 18.

[7] Marvin Minsky, *The Society of Mind* (Nueva York: Simon y Schuster, 1986) [Traducción española: *La Sociedad de la Mente*, Buenos Aires, Galápago, 1986].

[8] Ibid.

[9] Ibid.

[10] Edward A. Feigenbaum y Pamela McCorduck, *The Fifth Generation: Japan's Computer Challenge to the World* (Reading, Mass.: Addison-Wesley, 1983; edición en rústica, Nueva York: Signet, 1984) [Traducción española: *La Quinta Generación*, Barcelona, Sudamericana/Planeta, 1985].

[11] Roger Schank con Peter Childers, *The Cognitive Computer* (Reading, Mass.: Addison-Wesley, 1984).

[12] Harold Cohen, «What is an Image?», *Proceedings of the Sixth International Joint Conference on Artificial Intelligence*, Tokio (1979).

[13] Conversación privada con Harold Cohen.

[14] Pamela McCorduck, *The Universal Machine: Confessions of a Technological Optimist* (Nueva York: McGraw-Hill, 1985; edición en rústica, Nueva York: Harcourt Brace Jovanovich, 1986).

[15] Ibid.

[16] Correspondencia privada con Harold Cohen.

[17] Véase, por ejemplo, Arthur Danto, *The Transfiguration of the Commonplace* (Cambridge: Harvard University Press, 1981).

5
Redes neuronales e inteligencia artificial

Jack D. Cowan y David H. Sharp

Las redes neuronales son agregados de células nerviosas interconectadas, o neuronas. El cerebro humano, por ejemplo, es una red neuronal que comprende alrededor de diez mil millones de neuronas interconectadas. De alguna manera, esa red aprende y recuerda, piensa y siente. Es el sustrato de la conducta y la corporización de la mente. En el pasado medio siglo se han hecho intentos para modelizar las formas en que trabajan las redes neuronales, particularmente aquellas involucradas en la visión y el movimiento. En este artículo, sin embargo, nos concentraremos en un problema un tanto más abstracto, pero fundamental: la representación de sucesos externos dentro de las redes neuronales. Para nuestra concepción, este problema es central a toda comprensión de la conducta inteligente en mentes o máquinas. Concluiremos discutiendo las redes neuronales en relación con los estudios contemporáneos en inteligencia artificial.

Introducción

Las neuronas son células vivientes capaces de recibir y transmitir señales electroquímicas de manera altamente especializada. Sus complejidades se pueden simular adecuadamente sólo mediante intrincados chips de computadora; se necesitan redes que incluyan muchos de esos chips para simular incluso los procesos

---

Jack D. Cowan. Profesor de matemáticas aplicadas y biología teórica en el departamento de matemáticas de la Universidad de Chicago.

David H Sharp. Físico teórico en la división de teoría del Laboratorio Nacional de Los Alamos.

más sencillos que se piensa que ocurren en el cerebro. La modelización de redes neuronales, sin embargo, comenzó mucho antes de que esas complejidades fueran evidentes. Quizá la contribución principal fue un ensayo de Warren S. McCulloch y Walter H. Pitts publicado en 1943.[1] En este artículo McCulloch y Pitts aplicaron lógica simbólica al problema de describir lo que las redes neuronales pueden hacer. En efecto, probaron que todos los procesos susceptibles de describirse mediante un número finito de expresiones simbólicas (p.ej. la aritmética simple; la clasificación, almacenamiento y recuperación con conjuntos finitos de datos; la aplicación recursiva de reglas lógicas) puede encarnarse en redes de lo que ellos llamaron neuronas «formales». La fig. 1 muestra varios ejemplos de las neuronas de McCulloch y Pitts.

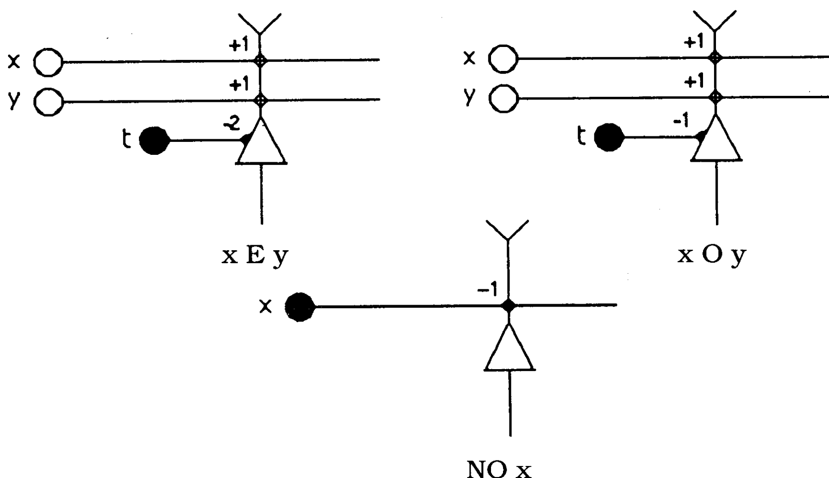


Figura 1. Neuronas formales de McCulloch-Pitts. Cada unidad se activa si y sólo si su excitación total alcanza o excede el valor cero. Por ejemplo, la primera unidad se activa si y sólo si ambas unidades x e y son activadas, pues sólo entonces la excitación total, $(+1)x+(+1)\,y$ balancea el umbral de -2 definido por la unidad de umbral, t, toda vez que x e y sean iguales a +1 (activadas). La unidad t siempre está activa. Los números (±1) y otros más mostrados arriba se llaman pesos. Los pesos positivos denotan sinapsis excitativas; los pesos negativos, sinapsis inhibitorias. Del mismo modo, los círculos abiertos denotan neuronas excitatorias; los círculos rellenos, inhibitorias.

Redes de McCulloch-Pitts

Las redes de neuronas formales o redes de McCulloch-Pitts, como se las llama ahora, son representaciones extremadamente simplificadas de la cosa real. Por ejemplo, son sincrónicas: la conmutación sólo ocurre a intervalos regulares y discretos. De este modo, las neuronas formales son sólo conmutadores lógicos simples, a diferencia de las neuronas reales. A despecho de estas simplificaciones, las redes de McCulloch-Pitts son importantes porque pueden dar cuerpo a cualquier operación o proceso que se pueda describir en términos lógicos. Donald A. Mackay ha expresado esta capacidad como sigue: si usted afirma que hay cierto proceso que una computadora no pueda realizar, y si puede describir en palabras exactamente qué es lo que constituye el proceso, entonces existe por lo menos una red de McCulloch-Pitts que puede encarnar y llevar adelante el proceso.[2] McCulloch y Pitts probaron de este modo que las redes neuronales formales, si se implementan con almacenamientos de memoria indefinidamente grandes, son equivalentes a la clase de máquinas de computación que Alan M. Turing ha demostrado que son computacionalmente universales.[3]

Computación confiable con neuronas no confiables

Las redes de McCulloch-Pitts fueron los primeros ejemplos de redes neuronales modélicas diseñadas para ejecutar tareas lógicas específicas. Pero, ¿qué pasa si esas redes cada tanto funcionan mal o si se dañan? Este problema atrajo a uno de los matemáticos más eminentes de este siglo, un pionero en el desarrollo de las computadoras digitales, John von Neumann. Al introducir la redundancia —utilizando muchas neuronas para hacer el trabajo de una sola—[4] von Neumann resolvió el problema de hacer que el funcionamiento de las redes de McCulloch-Pitts fuera confiable. En esas redes un bit de información (la elección entre uno y cero) se señala mediante la activación sincrónica de muchas neuronas, más que por medio de una activación todo-o-nada de una neurona formal: se obtiene uno cuando más de la mitad está activada, y cero en caso contrario. Von Neumann probó que las redes de McCulloch-Pitts redundantes operando en esta forma se pueden diseñar para desarrollar cálculos aritméticos con

alta confiabilidad. El trabajo subsiguiente de Shmuel Winograd y Jack D. Cowan proporcionó formas más eficientes para construir redes neuronales redundantes altamente confiables, al costo de requerir neuronas similares a microchips más complicadas, cada una con muchos contactos, para implementar las funciones lógicas requeridas.[5] La construcción de Winograd-Cowan fue notable porque utilizó una representación distribuida de la información: un bit de información se representaba en forma redundante mediante varias neuronas, como en la red de von Neumann, pero además cada neurona representaba parcialmente muchos bits.

Estas soluciones al problema de la confiabilidad proporcionaron una comprensión de la forma en que las redes neuronales en el cerebro pueden funcionar de manera confiable a pesar del daño. Desde los estudios neurológicos de John Hughlings Jackson de pacientes con daño cerebral[6] y la demostración de Karl S. Lashley de las habilidades cognitivas disminuidas en ratas con daño cerebral,[7] se ha tornado evidente que aunque las diferentes regiones del cerebro están especializadas para diversas funciones, la escala de esa localización de funciones no necesita extenderse a neuronas singulares. En términos del análisis de von Neumann-Winograd-Cowan, la representación de un bit de información no necesita ser unaria, sino que puede ser redundante o aun distribuida. Ha habido mucho debate sobre este punto. Lashley, por ejemplo, propuso que las diversas regiones del cerebro son equipotentes con respecto a la función[8] (cualquier región puede incorporar una tarea determinada), la antítesis misma de la localización regional. A la inversa, Horace B. Barlow ha afirmado que cuanto más uno se mueve de las regiones periféricas del cerebro a las regiones centrales, más se reduce el nivel de redundancia del funcionamiento cerebral.[9] El movimiento hacia la región central culmina en una representación unaria de la información profundamente en el cerebro. En la terminología actual, hablamos de neuronas «abuelas», que se activan supuestamente sólo cuando se percibe una abuela.

Conjuntos de células y sinapsis de Hebb

La noción de Lashley de la equipotencialidad de las regiones del cerebro se refleja en la obra de Donald O. Hebb.[10] En 1949 Hebb propuso que la conectividad del cerebro está cambiando conti-

nuamente a medida que un organismo aprende tareas funcionales diferentes y que se crean conjuntos de células debido a esos cambios. Hebb siguió una sugerencia temprana de Santiago Ramón y Cajal, postulando que la activación repetida de una neurona por otra a través de un contacto particular o sinapsis aumenta su conductividad, de modo que grupos de células débilmente conectadas, si se activan sincrónicamente, tienden a organizarse en conjuntos más fuertemente conectados. Aquí, de nuevo, la representación de un bit de información es distribuida. La propuesta de Hebb ha probado ser muy influyente. A despecho de la falta de evidencia en soporte de las ideas de Hebb, la teoría del conjunto de células disparó muchas investigaciones del aprendizaje en redes neuronales y de la forma en que la actividad neuronal sincronizada se genera y propaga.

Reconocimiento de patrones, aprendizaje y memoria

La propuesta de Hebb de la modificación sináptica durante el aprendizaje impulsó muchos trabajos sobre redes neuronales adaptativas, las que pueden aprender a ejecutar tareas específicas. Los trabajos iniciales sobre estas redes se desarrollaron en la década de 1950; Albert M. Uttley demostró que las redes neuronales con conexiones modificables a la manera de Hebb pueden por cierto aprender a clasificar conjuntos simples de patrones binarios (111010100, 101110101, etc) en clases equivalentes (p.ej., todos los que comienzan con 101).[11]

El problema de la clasificación o reconocimiento de patrones es central para cualquier teoría de la conducta inteligente, sea en animales o en máquinas. Pitts y McCulloch estuvieron entre los primeros investigadores que abordaron este problema.[12] Ellos notaron que los animales necesitan reconocer muchas versiones diferentes del mismo patrón, lo mismo que nosotros necesitamos leer muchas versiones diferentes del mismo texto: manuscrito, impreso en diferentes medidas, tipo de letra y color, visto bajo diferentes clases de iluminación. En efecto, la necesidad es reconocer no sólo un ejemplo de un patrón sino todos los ejemplos. Pitts y McCulloch construyeron dos redes neuronales, cada una de las cuales resolvió el problema en parte. La primera red intenta encontrar propiedades invariantes en un patrón dado (es decir,

propiedades comunes a todas las variaciones posibles del patrón). La segunda red transforma cualquier variante presentada externamente en una representación estándar. Pitts y McCulloch dieron entonces un paso temerario: propusieron que las redes neuronales de los córtices auditivo y visual representaban la primera solución y que la red neuronal en el colículo superior (que participa en el control del movimiento de los ojos) representaba la segunda. Se suponía que ambos córtices contenían un mecanismo que barría o muestreaba todas las variantes de un patrón a una frecuencia correspondiente al bien conocido ritmo alfa del córtex,[13] aproximadamente diez ciclos por segundo. McCulloch y Mackay desarrollaron más tarde experimentos que comprobaron que la hipótesis del barrido era falsa.[14]

Perceptrones

Unos diez años después de la publicación del artículo de Pitts y McCulloch, Frank Rosenblatt introdujo una estrategia importante en el problema del reconocimiento de patrones; Rosenblatt

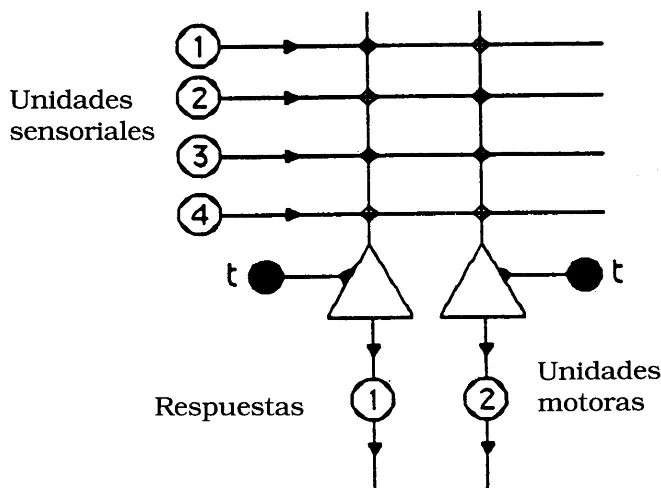


Figura 2. Un perceptrón elemental, es decir, una red neuronal adaptativa McCulloch-Pitts con pesos sinápticos modificados que se cambian si generan respuestas incorrectas.

demostró la forma en que las redes de McCulloch-Pitts con conexiones modificables podían ser «entrenadas» para clasificar ciertos patrones como iguales o distintos.[15] Rosenblatt llamó «perceptrones» a esas redes, y nosotros usaremos ese término en lo que sigue. La figura 2 muestra la arquitectura de un típico perceptrón elemental. Consiste de un conjunto de unidades «sensoriales» conectadas a otro conjunto de unidades «motoras», a través de un simple nivel de neuronas de McCulloch-Pitts (a las que nos referiremos como unidades M-P). Inicialmente, las fuerzas o pesos de todos los contactos o sinapsis en la red se ajustan a valores arbitrarios, de manera que la estimulación genere respuestas arbitrarias. Obtener una respuesta deseada de la red requiere ajustar todos esos pesos sinápticos. Rosenblatt encontró la forma de obtener la respuesta deseada mediante el siguiente proceso de entrenamiento: primero, anotar las respuestas de una unidad M-P a un estímulo determinado. Algunas de estas respuestas serán correctas (es decir, serán las respuestas deseadas); otras serán incorrectas. Luego, ajustar los pesos de las unidades como sigue: no hacer ningún ajuste si la respuesta es correcta. Si es incorrecta, en cambio, aumentar los pesos de todas las sinapsis si la unidad debía estar activada pero no lo está, o disminuirlos en caso contrario. Hacer lo mismo para todos los patrones deseados posibles de estímulo-respuesta. Se puede demostrar que después de sólo un número finito de presentaciones de patrones de estímulo-respuesta, los pesos convergen a un conjunto de valores representando cualquier computación o clasificación que corresponda a esos patrones.[16]

Adalines

Poco después de las primeras publicaciones de Rosenblatt apareció una variante de perceptrón estrechamente relacionada, inventada por Bernard Widrow y M.E. Hoff. La llamaron adaline (por *adaptive linear neuron*, neurona lineal adaptativa).[17] La única diferencia entre los perceptrones y los adalines radica en el procedimiento de entrenamiento. En el adaline la excitación liberada a una unidad M-P determinada se sustrae de la actividad deseada (definida +1 para activación y -1 para no activación, en lugar de 1 y 0). Llamemos al resultado *d*. El peso de una sinapsis activada se incrementa si *d* es positivo, se disminuye si *d* es negativo. A la

inversa, el peso de una sinapsis desactivada se aumenta si *d* es negativo, disminuye si es positivo. Esta regla corresponde estrechamente a la del perceptrón, pues si una unidad M-P no se activa mediante una unidad sensorial determinada cuando debería hacerlo, el peso de todas las sinapsis relevantes aumenta, y si lo inverso es verdad, decrece.

Limitaciones de perceptrones y adalines elementales

Hay límites para el rendimiento de los perceptrones y adalines elementales. Seymour A. Papert y Marvin L. Minsky probaron que los perceptrones elementales no pueden distinguir entre patrones tan simples como T y C.[18] La dificultad radica en la naturaleza de las unidades M-P. Como dijimos antes, unidades singulares de este tipo sólo pueden computar funciones lógicas tan simples como *x* AND *y*, *x* OR *y*, NOT *x*, *x* AND NOT *y*, etcétera. Sin embargo, la función *x* OR ELSE *y* y su negación NOT(*x* OR ELSE *y*), requieren varias unidades M-P. La razón es simplemente que *x* OR ELSE *y* es lo mismo que (*x* AND NOT *y*) OR (*y* AND NOT *x*). Esta

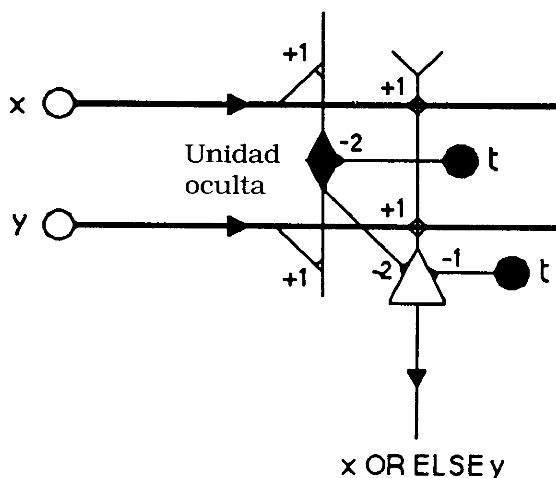


Figura 3. Una red M-P que implementa la función *x* OR ELSE *y*. La red comprende dos unidades M-P, una de las cuales es una unidad oculta, y dos unidades de umbral. Los pesos de las unidades ocultas no se pueden modificar correctamente con los procedimientos antes descritos.

situación es infortunada debido a que la función NOT(*x* OR ELSE *y*) es computacionalmente universal en el sentido de Turing: toda otra función se puede expresar como una cadena de NOT(*x* OR ELSE *y*)s. La figura 3 muestra la arquitectura de la red M-P más simple que implementa la función *x* OR ELSE *y*. Comprende dos unidades M-P y dos unidades *t*. Una de las unidades M-P está enteramente dentro de la red; su salida conduce sólo a la unidad M-P, no a la unidad motora. En la terminología actual se dice que el interior está «oculto», y lo que Papert y Minsky probaron es que un perceptrón o un adaline elemental, que consiste en un solo nivel de unidades M-P, no es computacionalmente universal, aunque tenga conexiones modificables. Por añadidura, conjeturaron que no se podían entrenar unidades ocultas en perceptrones de múltiples niveles; en otras palabras, que el problema de asignar crédito a unidades ocultas es insoluble.

Resultó después que las limitaciones de los perceptrones y adalines simples podían superarse. De hecho, en 1961 Rosenblatt introdujo un procedimiento de entrenamiento que casi resolvía el problema.[19] Pero a pesar de esa innovación, los procedimientos de entrenamiento exitosos no aparecieron hasta 1985. Los describimos en una sección más adelante.

Memoria asociativa

Otro rasgo notable del perceptrón es que su memoria del trabajo aprendido se distribuye sobre las conexiones modificadas durante la fase de entrenamiento y es por lo tanto poco probable que se perturbe por daño. A estos respectos, responde a algunas de las preocupaciones de Lashley sobre la memoria humana. Sin embargo, hay un importante aspecto de la memoria humana que los perceptrones no pueden afrontar directamente: a saber, la memoria humana parece ser asociativa además de distribuida. Lo que es común a dos recuerdos diferentes de alguna manera los liga, de modo que uno puede evocar al otro si hay entre ellos suficiente superposición.

Las redes neuronales con memoria asociativa se han estudiado extensivamente desde mediados de la década de 1950, comenzando con los trabajos de Wilfrid K. Taylor.[20] La figura 4 muestra la estructura de la red original de Taylor. Consiste en un nivel de unidades asociativas emparedadas entre conjuntos de unidades

sensorias y motoras. Es similar en estructura a un perceptrón de tres capas, excepto que todos los pesos de los contactos en la red son modificables, y las unidades no son neuronas M-P sino dis-positivos analógicos. (Considérese la diferencia entre una bobina reductora y un conmutador de luz. Con la bobina se puede cam-biar el nivel de iluminación en forma gradual, mientras que con un conmutador es todo o nada. Los dispositivos analógicos operan como bobinas más que como conmutadores.) El procedimiento de entrenamiento, también diferente del entrenamiento del perceptrón, es simplemente la regla de Hebb: los pesos sinápticos ac-tivados aumentan si se activan las unidades de destino. Esos cam-bios se habían observado en el tejido cerebral.[21] La red aprende a asociar patrones sensoriales diferentes mediante la presentación repetida de pares de patrones, uno de los cuales inicialmente suscita una respuesta motora. Con el tiempo, los otros patrones disparan la respuesta. De este modo, las redes de Taylor exhiben un simple condicionamiento pavloviano,[22] y la memoria asociada se almacena en forma distribuida en el patrón de pesos.

Posteriormente Taylor construyó una red más elaborada en la que las unidades motoras reconectaban con unidades sensoriales

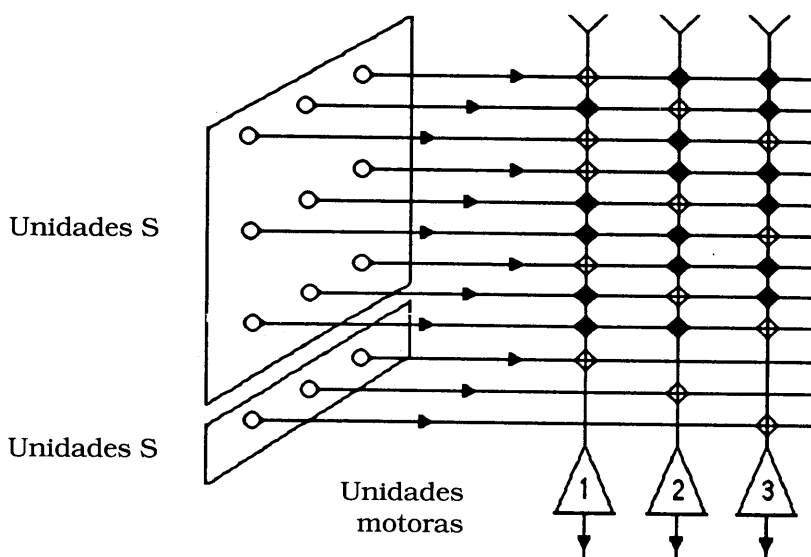


Figura 4. Una red de Taylor. Esta red utiliza unidades analógicas con pesos modificables y se puede entrenar para asociar diferentes conjuntos de patrones de estímulo.

y entre sí. Esa red es capaz de formar asociaciones con estímulos apareados de una manera más confiable y controlable que en la red anterior, y es también capaz de discriminar patrones al estilo de los perceptrones y adalines. Taylor sugería que las áreas de asociación de la corteza cerebral y el tálamo contenían esas redes.[23]

Poco después de esto, Karl Steinbuch introdujo una red muy parecida, la «matriz de aprendizaje».[24] Consiste de un conjunto de conmutadores interpuestos entre unidades sensorias y motoras. Como en el esquema de Taylor, la red aprende a asociar patrones sensoriales con patrones motores. La memoria asociada se almacena en un patrón de conmutadores abiertos y cerrados. Las matrices de aprendizaje poseen una estructura matemática particularmente simple, y su comportamiento se puede analizar rápidamente. Siguiendo a Steinbuch, pero en la mayoría de los casos en forma más bien independiente, muchos otros desarrollaron redes similares, por ejemplo James A. Anderson,[25] David J. Willshaw, O. Peter Buneman y H. Christopher Longuet-Higgins,[26] David Marr[27] y Teuvo Kohonen,[28] quienes descubrieron que las redes asociativas son también direccionables por contenido (es decir, estimulando una de esas redes con algún fragmento de una memoria asociada, se elicitará la respuesta completa). De esta manera, la red se puede direccionar con el contenido parcial de una memoria antes que especificando su ubicación. Debido a esta propiedad, las redes asociativas se conocen ahora como memorias asociativas direccionables por contenido (ACAMs). El trabajo de Marr sobre esta propiedad es particularmente interesante por cuanto se formula como una teoría de la forma en que el cerebelo permite a los animales realizar movimientos voluntarios delicados y precisos, y de cómo la memoria se puede albergar momentáneamente en el hipocampo.

Una teoría del cerebelo

El cerebelo («pequeño cerebro») está presente en todos los vertebrados. Tiene más o menos tantas neuronas como el cerebro, y se piensa que es el órgano que controla los movimientos voluntarios. Comparado con el cerebro, el cerebelo tiene una arquitectura sorprendentemente regular y simple. Esta arquitectura se reveló en la década de 1960,[29] pero su función precisa y su modo de operación están por descubrirse. Marr introdujo la idea de que el cerebelo es

un ACAM que es entrenado por el cerebro para controlar la ejecución de secuencias de movimientos voluntarios.[30] En la teoría de Marr (véase fig. 5), se asigna una función a cada uno de los cinco tipos de neurona que comprenden la red cerebelar. Se sabe que los ACAMs trabajan más eficientemente cuando los patrones almacenados no están correlacionados entre sí. Marr asigna a las células granulares la tarea de descorrelacionar los patrones de actividad que llegan por las fibras musgosas. Los patrones resultantes se guardan en el cerebelo por la vía de las sinapsis de Hebb, entre las células granulares y las células de Purkinje y bajo el control de la activación de las fibras en pendiente, exactamente como se guardan los patrones en una matriz de aprendizaje. Es fácil sobrecargar a los ACAMs con demasiados patrones de fibras musgosas. Se supone que las células de Golgi impiden la sobrecarga elevando los umbrales de las células granulares. Dado que las células de Golgi están guiadas por la actividad de las fibras musgosas, cuanto más activas están, más se inhiben y por lo tanto levantan los umbrales de las células granulares. De este modo, las células de Golgi actúan como los controles automáticos de volumen de radios y televisores. Para

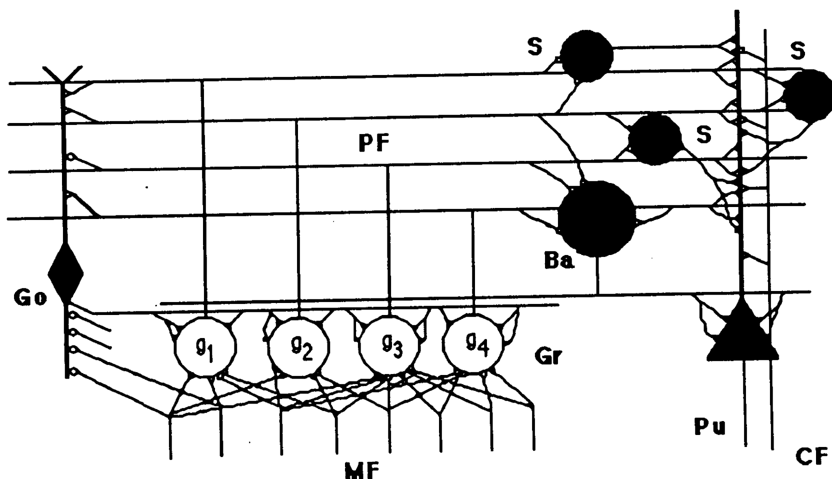


Figura 5. La teoría del cerebelo de Marr. Las células granulares (g) son las únicas células excitatorias en el cerebelo; todas las demás son inhibitorias. Las células de Golgi (Go) controlan los umbrales de las células granulares; las células de cesta (Ba) y las estrelladas (S) controlan los umbrales de las células de Purkinje (Pu). La red se entrena en el estilo ACAM estándar para asociar los patrones de las fibras musgosas (MF) y de las células en pendiente (CF).

recuperar correctamente los patrones del almacenamiento, los umbrales de activación de las células de Purkinje se deben fijar lo suficientemente altos como para suprimir los patrones no deseados. El resultado de todo esto es que la activación de las fibras en pendiente entrena a la red cerebelar para que responda apropiadamente a los patrones de activación de las fibras musgosas. Se postula que este entrenamiento corresponde al aprendizaje para ejecutar secuencias complicadas de movimientos voluntarios (p.ej., conducir, volar, tocar el piano).[31]

La teoría de Marr es notable porque quizá por primera vez se asignaba una función específica a cada neurona en una parte del cerebro. Esta teoría fue modificada ligeramente por James S. Albus, quien señaló que, dado que las células de Purkinje inhiben células en el núcleo cerebelar, es más plausible que las sinapsis de las células granulares y de Purkinje se debiliten y no que se refuercen en caso de activación coincidente (es decir, el entrenamiento debilita la inhibición de las células del núcleo cerebelar por las células de Purkinje).[32] Desde la publicación de los ensayos de Marr-Albus, se han hecho numerosos intentos de probar la teoría.[33] Hay que admitir que su validez o invalidez no ha sido definitivamente determinada, aunque algunos resultados experimentales recientes parecen dar soporte a la versión de Albus de la teoría.

Una teoría del hipocampo

Marr aplicó un análisis similar a otra parte del cerebro, el hipocampo,[34] así llamado porque su forma se parece a la de un caballito de mar. El hipocampo se encuentra en el lóbulo temporal del cerebro y se piensa que es la región donde se forma la memoria a corto plazo o memoria de trabajo, particularmente cuando está relacionada con los aspectos espaciales del entorno del animal.[35]

Al igual que el cerebelo, el hipocampo tiene una estructura particularmente regular. Su región principal, el *cornu ammonis* (CA), comprende una simple página de neuronas de salida, las así llamadas células piramidales, junto con «interneuronas» subordinadas, en su mayoría neuronas esteladas. En la teoría de Marr las células piramidales son análogas a las células de Purkinje en el cerebelo, y las diversas interneuronas son análogas a las células granulares, de Golgi, en cesta y esteladas. De este modo el hipocam-

po, como el cerebelo, se modeliza como un ACAM. Sin embargo, hay importantes diferencias entre las dos estructuras. De acuerdo con Marr, el hipocampo tiene que aprender a formar sus propias clasificaciones internas de los numerosos patrones de entrada que tiene que almacenar. Por lo tanto, necesita células granulares con sinapsis modificables tanto como células piramidales. De hecho, Marr demostró que al menos se necesitan dos páginas de células granulares modificables para una operación confiable. Por añadidura, el hipocampo, dada sólo una fracción pequeña de las pistas relevantes, tiene que ser capaz de recuperar los patrones. Sólo puede hacerlo si las células piramidales del CA están interconectadas vía sinapsis de Hebb excitatorias y modificables, de modo que toda la página de las pirámides del CA actúen cooperativamente.

Una teoría del neocórtex cerebral

La teoría del hipocampo de Marr es en realidad una especialización de su teoría más general sobre la función del neocórtex cerebral.[36] El neocórtex, que contiene la mayoría de las neuronas en el cerebro, constituye la mayor parte del conjunto cerebral. Marr postuló que la función primordial de las redes neocorticales es la de formar representaciones internas de clases y subclases de objetos, utilizando procedimientos similares a los del perceptrón. Al proporcionar a las fibras en pendiente el papel de activar y guiar la formación de nuevas neuronas clasificatorias, esta teoría difiere de la teoría del hipocampo. Hasta hoy, ninguno de estos modelos ha sido definitivamente probado. Sin embargo, se ha acumulado considerable evidencia de que muchas de las sinapsis excitatorias de las células piramidales del hipocampo se pueden reforzar durante tiempos largos (de segundos a minutos) mediante una estimulación presináptica adecuada. Este efecto, llamado potenciación a largo plazo (LTP),[37] es consistente con las teorías corticales de Marr.

Desarrollos actuales

Después del trabajo de Marr ha habido un largo hiato durante el cual se progresó muy poco en la forma de entrenar redes neuronales para representar información. Se ha avanzado bastante en el

estudio de la forma en que se desarrolla el cerebro[38] y en neurodinámica,[39] la generación y propagación de actividad neuronal sincronizada. No fue hasta comienzos de la década de 1980, sin embargo, que se progresó realmente en los problemas anticipados por Rosenblatt y Marr. Estas nuevas investigaciones se clasifican generalmente bajo el rubro de conexionismo, que denota la vieja noción de que la información está almacenada en el cerebro como un patrón de pesos sinápticos definidos durante el aprendizaje. Esta idea ha estado rondando casi desde que Ramón y Cajal descubrió las neuronas[40] y, como ya hemos señalado, fue elaborada por Hebb en 1949. Constituye la base de casi todo el trabajo sobre perceptrones y adalines desde entonces. De aquí en más utilizaremos el término *neoconexionismo* para referirnos a los trabajos actuales.

Redes de Hopfield

Como primer ejemplo del neoconexionismo describiremos el trabajo de John J. Hopfield,[41] quien demostró la analogía formal entre una red de elementos similares a las neuronas con conexiones simétricas,[42] ahora llamada red de Hopfield, y un material descubierto en la pasada década, llamado *spin glass*.[43] Los orígenes de este trabajo se encuentran en un ensayo sumamente agudo publicado en 1954 por el neuroanatomista Brian G. Cragg y el físico Nevill V. Temperley.[44] Cragg y Temperley observaron que así como las neuronas pueden estar activadas o inactivas, los átomos en un conjunto o grilla pueden estar en uno de dos estados: con los spins apuntando «para arriba» o con los spins apuntando «para abajo» (véase fig. 6). Además, así como las neuronas excitan o inhiben a otras, del mismo modo los átomos ejercen sobre sus vecinos fuerzas que tienden a colocar los spins en la misma dirección o en la dirección opuesta. Las propiedades de las neuronas en una red densamente conectada son probablemente similares a las de los átomos (o aleaciones binarias) en una red cristalina. Los cristales, las aleaciones y otros conjuntos atómicos pueden mostrar distintos tipos de orden y desorden. Estos van desde un orden a corto plazo, en el cual, en promedio, cada spin hacia arriba está rodeado por spins hacia abajo, hasta un orden a largo plazo, en el cual, en promedio, los spins hacia arriba persisten —digamos— en una de cada tres celdas de la red

en una dirección determinada. Los sistemas de spins que muestran distintos tipos de orden proporcionan buenos modelos de las propiedades de los materiales magnéticos. Por ejemplo, un ferromagneto, que consiste en átomos que tienden a forzar a los demás a colocar el spin en la misma dirección, posee un orden a largo plazo; un antiferromagneto, que consiste en átomos que tienden a forzar los spins de los demás en dirección opuesta, poseen también un orden a largo plazo; por otra parte, un paramagneto, que consiste en átomos con el spin tanto para arriba como para abajo en patrones al azar, está desordenado. Es posible que las redes neuronales exhiban propiedades análogas. Cragg y Temperley por lo tanto sugirieron: a) que los patrones del dominio que son un rasgo presente de los ferromagnetos, y que comprenden placas de spins hacia arriba o hacia abajo, deben aparecer en las redes neuronales como placas de neuronas excitadas o inactivas, y b) que los patrones del dominio neuronal, una vez activados por estímulos externos, serían estables frente a actividad espontánea al azar y podrían constituir por ello una memoria del estímulo.[45] Es interesante señalar que veinte años más tarde William A. Little, mediante el análisis matemático de un sistema de spin en malla, llegó virtualmente a las mismas conclusiones de Cragg y Temperley concernientes a la existencia de estados neuronales persistentes.[46]

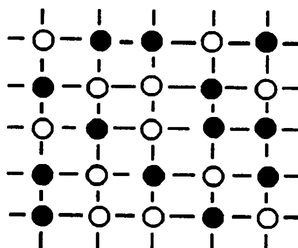


Figura 6. Sistema de malla de spins. Los spins «hacia arriba» y «hacia abajo» se ordenan en una malla cuadrangular. Cada spin interactúa con sus vecinos más cercanos para asumir una configuración estable.

En 1975 David Sherrington y Scott Kirkpatrick descubrieron un nuevo material magnético consistente en una mezcla al azar de spins interactuando ferromagnéticamente y antiferromagnéticamente y que no exhibía magnetismo de red.[47] Llamaron a este material *spin glass*. Los cristales de spin poseen propiedades

interesantes, una de las cuales es la capacidad de almacenar muchos patrones desordenados de spin diferentes. Las redes de Hopfield poseen propiedades similares, pero no son redes neuronales, dado que cada elemento debe tanto excitar como inhibir a los elementos vecinos (véase fig. 7). Sin embargo, son de interés como redes neuronales artificiales, particularmente para memoria de almacenamiento.

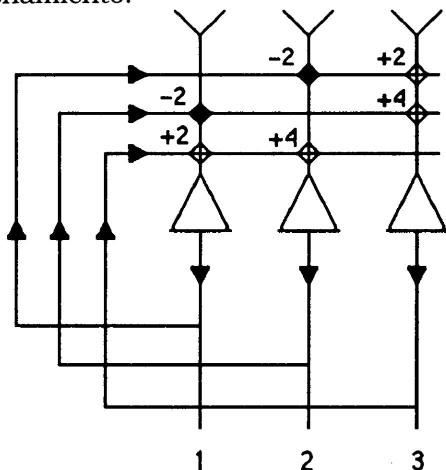


Figura 7. Una red de Hopfield con conexiones simétricas. Las unidades 1 y 2 excitan una vecina e inhiben a otra.

Hopfield reconoció la analogía formal entre una red de elementos similares a las neuronas con pesos de conexión simétricos al azar y una malla de spins y, utilizando la regla para la modificación de los pesos sinápticos postulada por Hebb, demostró que los pesos se pueden modificar de manera tal que se estabilice la actividad de la red.[48] Dados esos pesos, cualquier configuración inicial de elementos activos e inactivos evolucionará hacia una configuración estable. De esta manera, las configuraciones estables se pueden usar para almacenar confiablemente información. Las redes de Hopfield, de hecho, sirven como ACAMs confiables y son similares en muchos aspectos a los construidos por Taylor, Steinbuch, Marr y sus asociados.

Las redes de Hopfield representan un avance conceptual importante en la teoría de las redes neuronales. Aunque no son muy realistas como modelos de las redes nerviosas, el principio

que encarnan (almacenar información en configuraciones dinámicamente estables) es profundo. Este principio se origina en el trabajo de uno de los primeros cibernéticos, W. Ross Ashby, quien acuñó el término ultraestabilidad para describir la forma en que, según pensaba, los patrones de actividad del cerebro siempre tienden a configuraciones dinámicamente estables.[49] Este principio está implícito en el trabajo de muchos otros estudiosos.[50]

Computando con redes de Hopfield

Las redes de Hopfield han probado ser de interés para resolver problemas de optimización computacional. Una aplicación bien conocida de esas redes es el «problema del vendedor viajero», en el que un vendedor necesita visitar una vez cada una de unas cuantas ciudades, en un plan de trabajo que minimice la longitud del viaje. Este es un ejemplo de lo que se ha llamado un problema de optimización constreñida.[51] S. Kirkpatrick, C.D. Gelatt (h) y M. P. Vecchi han demostrado que las configuraciones de equilibrio asumidas por un cristal de spin proporcionan soluciones a este problema.[52] John J. Hopfield y David W. Tank han demostrado que ciertas redes de Hopfield también encuentran buenas solucio-

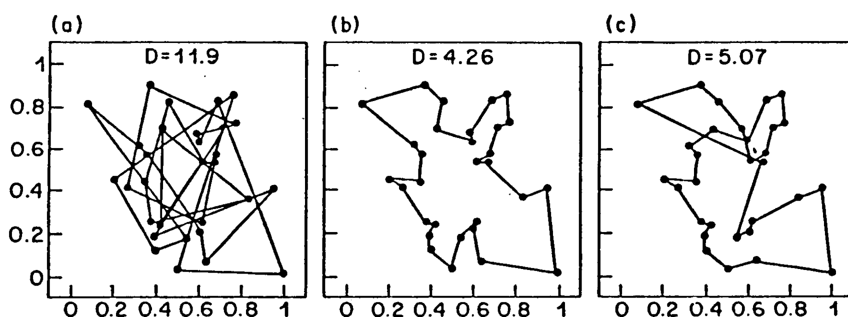


Figura 8. Comparación de procedimientos para resolver la estrategia del problema del vendedor viajero: (a) viaje al azar, de longitud total D=11,9; (b) viaje encontrado por el procedimiento Kernighan-Lin, D=4,26; (c) viaje encontrado por el procedimiento Hopfield-Tank, D=5,07. Se verá que el viaje de Hopfield-Tank es casi tan corto como el de Kernighan-Lin (copiado de J.J. Hopfield y D.W. Tank, «'Neural' Computation and Constraint Satisfaction Problems and the Traveling Salesman», *Biological Cybernetics* 55 [1985]:141.)

nes.[53] La figura 8 muestra una solución que involucra treinta ciudades. Aunque la red de Hopfield no encuentra el camino más corto, encuentra uno que se compara bastante favorablemente con la solución que se encuentra mediante el procedimiento de Kernighan-Lin,[54] uno de los mejores procedimientos para resolver problemas de optimización constreñidos. Recientemente Richard Durbin y David J. Willshaw han desarrollado otra red similar a las neuronas que encuentra caminos aún más cortos y que trabaja mejor en problemas más grandes.[55]

Máquinas de Boltzmann

Las redes de Hopfield sufren de un defecto en su habilidad para encontrar la mejor solución en problemas de optimización constreñida. Pueden quedar atrapadas en configuraciones metaestables. Para encontrar el mínimo global verdadero, la red debe hacer cambios configuracionales al azar de tiempo en tiempo, ganando así la capacidad de escapar de configuraciones metaestables. Esta es en esencia una versión bien conocida del procedimiento de Monte Carlo, introducida en el Laboratorio Nacional de Los Alamos en 1953 por N. Metropolis, A. Rosenbluth, M. Rosenbluth, M. Teller y E. Teller para encontrar estados estables.[56] En este procedimiento se computa el cambio producido por la conmutación rápida de uno de los spins de la red. Si la nueva configuración es más estable, se retiene. De otra manera, la configuración se rechaza (es decir, la conmutación es cancelada). Este procedimiento, aunque lento, encontrará con el tiempo la configuración más estable. Geoffrey E. Hinton y Terrence J. Sejnowski, de acuerdo con eso, utilizan el procedimiento de Monte Carlo para hallar configuraciones estables en redes de Hopfield,[57] repitiendo, en efecto, el uso del método de Monte Carlo por parte de Kirkpatrick, Gelatt y Vecchi en el problema del cristal de spin. Al hacerlo, descubrieron un proceso por el cual las redes resultantes, que ellos llamaron máquinas de Boltzmann, pueden modificar su conectividad de una manera que resuelve el problema de la asignación de crédito a las unidades ocultas.

La máquina de Boltzmann, una red adaptativa de Hopfield con unidades ocultas, implementa un procedimiento de Monte Carlo para encontrar la configuración estable de unidades activas e

inactivas; estas unidades no son simples neuronas M-P, que tienen respuestas todo o nada, sino dispositivos analógicos. Hinton y Sejnowski demostraron que con estos dispositivos se pueden alcanzar configuraciones estables si los pesos de contacto cambian mediante el siguiente procedimiento: definiendo *p* como la probabilidad promedio de que dos unidades estén simultáneamente activadas cuando las unidades *S* se activan mediante un patrón de estímulo que empalma las unidades motoras en algún patrón de activación, y siendo *p'* la probabilidad correspondiente cuando la máquina corre libremente en ausencia de estímulo, y *w* el peso del contacto entre las dos unidades, la regla por la cual cambia *w* es muy simple: si *p* es mayor que *p'*, incrementar *w*; si *p* es menor que *p'*, disminuirlo. (El lector debe comparar esta regla con las reglas de entrenamiento de los perceptrones y adalines en las páginas 108 y 109). Con tal regla en acción, las máquinas de Boltzmann son capaces de resolver una variedad de problemas de optimización constreñida.[58] Muchos de esos problemas aparecen en las estrategias computacionales referidas a la visión.

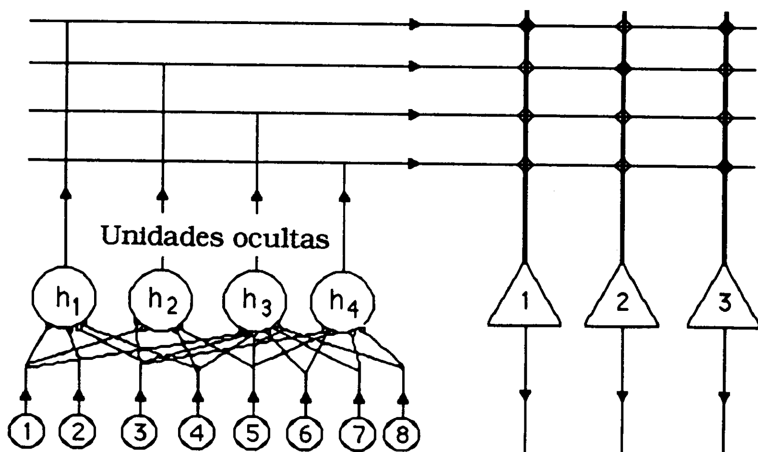
Representaciones del aprendizaje

La máquina de Boltzmann también proporciona una solución al problema de asignación de crédito a las unidades ocultas, si bien sólo en el caso especial de las redes de Hopfield adaptativas. La regla antes mencionada resuelve el problema de la asignación de crédito a las unidades ocultas sólo en términos de la información localmente disponible (es decir, cuando el cambio de peso del contacto entre dos unidades depende sólo de los patrones de activación). El proceso de aprendizaje de la máquina de Boltzmann es autoasociativo y no supervisado, depende sólo de correlaciones entre pares de unidades y crea en el conjunto de los pesos de conexión una representación distribuida de las correlaciones que existen en y entre los miembros del conjunto de patrones de estímulo. Para decirlo de otro modo, una máquina de Boltzmann puede formar una representación que eventualmente reproduce relaciones entre clases de sucesos en su ambiente.[59] Proporciona por lo tanto una posible solución al problema de Marr de cómo construir tales representaciones *ab initio* en las células granulares del hipocampo y el neocórtex. Más generalmente, proporciona una

vía para formar representaciones distribuidas de símbolos abstractos, y por lo tanto permite la investigación del razonamiento simbólico por medio de redes neuronales adaptativas.[60]

Propagación hacia atrás

La máquina de Boltzmann representa un avance considerable en el aprendizaje de máquina no supervisado. Sin embargo, debido a que la máquina utiliza una versión del procedimiento de Monte Carlo para encontrar configuraciones estables, el aprendizaje es muy lento. Además, la máquina de Boltzmann es una red de Hopfield con conexiones sólo simétricas. Estas limitaciones han sido superadas recientemente por David E. Rumelhart, Geoffrey Hinton, Robert J. Williams y otros en una exitosa implementación del procedimiento originalmente sugerido por Rosenblatt.[61] La figura 9 muestra la estructura de la red. Esencialmente se trata de un perceptrón de dos capas. Es también la arquitectura básica del modelo de Marr del cerebelo, sin interneu-



Unidades ocultas - Unidades sensoriales - Unidades motoras

Figura 9. Estructura de un perceptrón de múltiples capas. En contraste con un perceptrón elemental (cf. fig. 2), hay una capa de unidades ocultas, cuyos pesos sinápticos también son modificables. Esta estructura es, de hecho, un subconjunto del modelo de Marr del hipocampo, pero el procedimiento para cambiar los pesos es diferente.

ronas inhibitorias. Las reglas por las que se modifican los contactos difieren considerablemente de las del modelo de Marr, sin embargo, y derivan de las reglas del adaline que vimos antes. Recuérdese que en un adaline, los cambios de peso son proporcionales a las diferencias entre el patrón de activación deseado y la excitación total de la unidad. Rumelhart y otros han demostrado que para las unidades analógicas, una simple extensión de las reglas del adaline resuelve el problema de la asignación de crédito a las unidades ocultas.

Las computaciones concretas resultan afectadas en dos etapas. En la primera, la etapa hacia adelante, se estimula la red y se anotan las respuestas de la unidad motora. En la segunda, la etapa hacia atrás, se usan estas respuestas para ajustar los pesos de las mismas unidades motoras, y luego se ajustan los pesos de la unidad oculta: de allí la descripción del proceso como propagación hacia atrás. Es aquí como demuestra ser decisiva la diferencia entre los elementos analógicos utilizados por Rumelhart y otros y las simples unidades M-P utilizadas en los perceptrones y adalines elementales. El procedimiento de los adalines para modificar los pesos sinápticos de la unidad motora se puede extender a los pesos de la unidad oculta. Considérense los pesos de la fig. 10, que muestra una sección de la red ilustrada en la fig. 9. El cambio

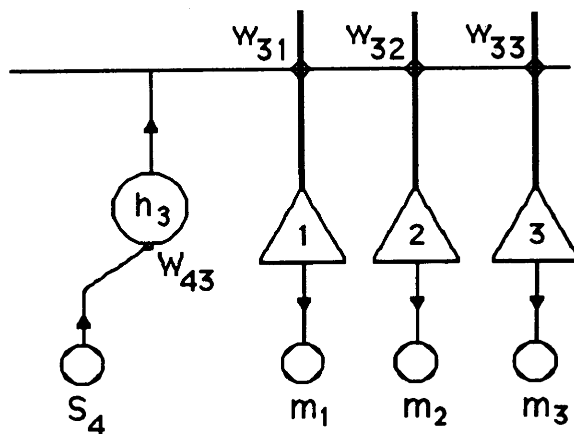


Figura 10. Cómo trabaja la propagación hacia atrás. Se estimula primero la red, en este caso por medio de la unidad S_4. Se obtienen entonces las respuestas motoras m_1, m_2 y m_3. Se ajustan luego los pesos w_{31}, w_{32} y w_{33}, y finalmente el peso W_{43}.

requerido en el peso de la unidad oculta W_{43} está relacionado con los pesos y los cambios de peso de todas las unidades «corriente abajo» de h_3. De este modo, dados los cambios en w_{31} etc., que están determinados por una ligera modificación del procedimiento del adaline que describimos antes (adaptado a unidades analógicas) se puede computar el cambio en W_{43}.

El procedimiento de Rumelhart et al. proporciona una solución al problema de la asignación de crédito y convierte a los perceptrones y adalines analógicos de capas múltiples en poderosas herramientas para investigar el aprendizaje supervisado en redes neuronales. En lo que sigue describimos unas pocas aplicaciones de la propagación hacia atrás.

El problema *x* OR ELSE *y*

Ya hemos señalado el fracaso de los perceptrones y adalines elementales para computar la función lógica *x* OR ELSE *y*. Esta función es verdadera si *y* sólo si *x* es verdad e *y* es falsa o viceversa, y es falsa de otro modo. La tabla de verdad correspondiente a *x* OR ELSE *y* se muestra en la página 126. Con propagación hacia atrás se obtiene la red ilustrada en la figura 11 después de un entrenamiento que comprende 558 pasadas por los cuatro patrones mostrados en la tabla de verdad. En este caso, tanto las unidades ocultas como la unidad motora poseen umbrales negativos y están activadas a menos que se las inhiba lo suficiente. La unidad oculta h_1 está activada si ninguna unidad *S* está activada, y cuando h_1 está activada desconecta la unidad motora. La unidad motora también se desactiva si ambas unidades *S* están activas. Esta red difiere un poco de la que se muestra en la fig. 3, en la que todos los sesgos de umbral son positivos; sin embargo, también computa la función lógica *x* OR ELSE *y*.

El problema T/C

Otro problema resuelto por la propagación hacia atrás es un problema geométrico: distinguir entre las letras *T* y *C* independientemente de la traslación y la rotación. La fig. 12 muestra los patrones y la fig. 13 la arquitectura de la red utilizada por

X	Y	X OR ELSE Y
+1	+1	-1
+1	-1	+1
-1	+1	+1
-1	-1	-1

Tabla de verdad para *x* OR ELSE *y*. +1 denota «verdad» y -1 «falso». De este modo, *x* OR ELSE *y* es falso cuando *x* e *y* son ambos verdad o ambos falsos. De otro modo *x* OR ELSE *y* es verdad.

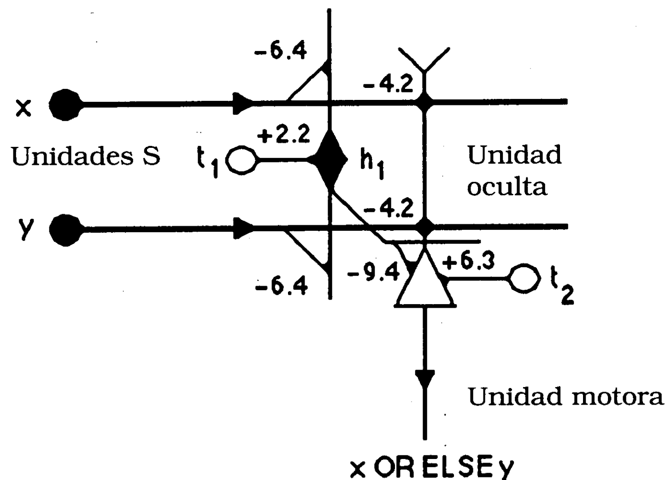


Figura 11. Red obtenida mediante propagación hacia atrás, que computa correctamente la función lógica *x* OR ELSE *y*. Las unidades t_1 y t_2 tienen umbral cero y están siempre activas (Reproducido de David E. Rumelhart, Geoffrey E. Hinton y Robert J. Williams, «Learning Internal Representations by Error Propagation», en el vol. 1 de *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, ed. David E. Rumelhart y James L. McClelland [Cambridge: MIT Press, 1986]).

Rumelhart et al. para resolver el problema.[62] Esta estructura es más o menos la misma que la que usaron Rosenblatt y otros para problemas similares. Es notable que las unidades ocultas estén conectadas sólo con pequeñas regiones de la página de unidades *S* (llamadas campos receptivos), mientras que la unidad motora

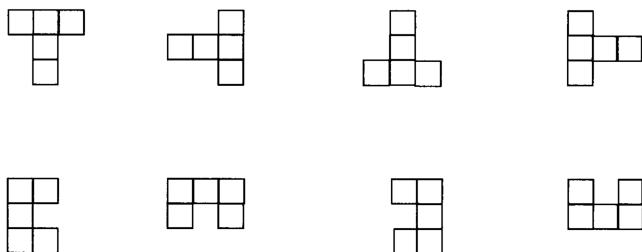


Figura 12. Patrones *T* y *C* a ser distinguidos por un perceptrón de propagación hacia atrás de tres capas (Reproducido de Rumelhart et al. [Véase fig. 11].)

está conectada con muchas unidades ocultas ampliamente separadas. Este patrón de conexión imita en cierta medida la arquitectura del cerebro visual.[63]

Después de unas 5.000 a 10.000 representaciones de los patrones *T* y *C*, junto con las respuestas apropiadas, la red aprende la tarea correspondiente. Al hacerlo, los campos receptivos de la unidad oculta devienen adaptados a la tarea en cierto número de formas, cuyo efecto es facilitar la distinción entre *Ts* y *Cs* por medio de la unidad motora final.

NETtalk

En una aplicación aún más sorprendente, Terrence Sejnowski y Charles R. Rosenberg entrenaron otra red similar para leer y hablar textos en inglés.[64] La red comprende 203 unidades *S* dispuestas en siete grupos de 29; 80 unidades *h* y 26 unidades motoras. En cada grupo de unidades *S*, 26 codifican una letra del alfabeto inglés y las 3 restantes codifican la puntuación y los límites entre palabras. El patrón de estímulo es entonces una hilera de 7 caracteres. Las unidades motoras codifican sonidos del habla, o fonemas, y también acentos y hiatos entre sílabas. La red fue entrenada utilizando propagación hacia atrás con cierto número de textos. Un experimento utilizó transcripciones fonéticas del habla continua informal de un niño. Se usaron aproximadamente 1.000 palabras del corpus, y después de unas 50.000 presentaciones, la red era capaz de leer y hablar con una exactitud del 95 por ciento. La fig. 14 muestra el campo receptivo de una de

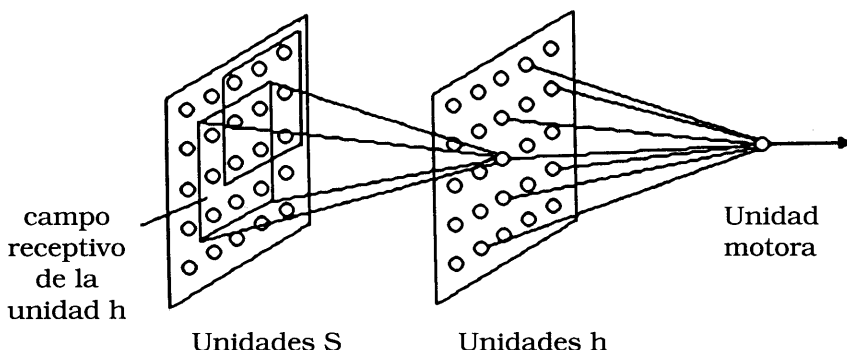


Figura 13. Estructura de un perceptrón de tres capas que puede distinguir Ts de Cs. La capa de entrada de las unidades S es una página bidimensional, como lo es la capa de unidades ocultas. El campo receptivo de cada unidad h comprende un cuadrado de 3 x 3 unidades S; es decir, se encuentra limitado por su superficie (cf. Minsky y Papert, *Perceptrons: An Introduction to Computational Geometry* [Cambridge: MIT Press, 1969]). Los pesos de las unidades ocultas se ajustan mediante propagación hacia atrás en el curso de unas 10.000 presentaciones de patrones de T y C y convergen en un conjunto que distingue los patrones T de los patrones C (Reproducido de Rumelhart et al. [Véase fig. 11]).

las unidades ocultas. Es evidente que esta unidad posee una representación distribuida de muchos atributos de las hileras de entrada.

Se presentaron luego a la red 439 continuaciones de palabras procedentes de textos del mismo niño (que contenía muchas palabras nuevas), que la red leyó y pronunció con una exactitud del 78 por ciento. Este es un ejemplo de generalización. La red de Sejnowski-Rosenberg, llamada NETtalk, generaliza bastante bien. De hecho, los perceptrones y los adalines de capas múltiples generalizan bastante bien a lo largo de toda una variedad de tareas, tal como Rosenblatt lo había anunciado varios años atrás.[65] Otra propiedad exhibida por NETtalk es la resistencia al daño. Una red NETtalk sustancialmente dañada puede aún leer y hablar con una adecuación de alrededor del 40 por ciento, y se recupera rápidamente con el entrenamiento. Cabe esperar tales propiedades en redes con representaciones distribuidas, tal como lo sugería la teoría de Winograd-Cowan. En general, las redes

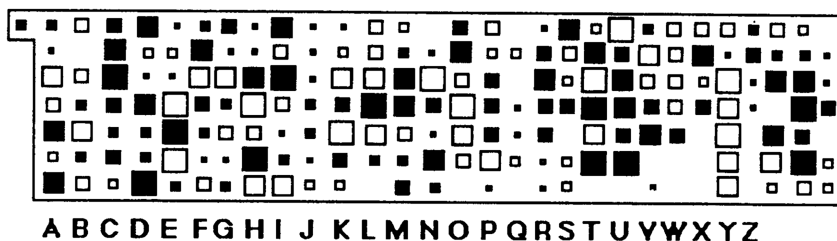


Figura 14. Campo receptivo de una unidad oculta en NETtalk. Comprende 203 unidades *S* más una unidad *h* para definir su umbral. Las unidades *S* se disponen en siete grupos de 29. Veintiséis de las 29 unidades *S* codifican letras del alfabeto, y 3 de ellas codifican signos de puntuación y espacios. De este modo, cada unidad oculta responde a una serie de siete caracteres de un modo específico definido por sus pesos. El área de cada cuadrado es proporcional al peso; los cuadrados abiertos corresponden a pesos positivos, los cuadrados negros a los pesos negativos. (Reproducido de T.J. Sejnowski y C.R. Rosenberg, *NETtalk, A Parallel Network that Learns to Read Aloud*, Technical Report JHU/EECS-86/01 [Baltimore, Md.: Johns Hopkins University, Electrical Engineering and Computer Science, 1986].)

NETtalk trabajan sorprendentemente bien, aunque tienen poca habilidad para tratar con ambigüedades sintácticas y semánticas. Puede esperarse que versiones mejor estructuradas de NETtalk se comporten mejor a este respecto.

Arboles familiares

Describimos como ejemplo final de la propagación hacia atrás una red que es entrenada para almacenar relaciones abstractas[66] (véase la información en los dos árboles familiares ilustrados en la fig. 15). La red comprende 24 unidades S, cada una de las cuales representa una persona; doce unidades *S* adicionales, cada una representa una relación; 30 unidades ocultas dispuestas en tres capas y 24 unidades motoras, cada una representa una persona. La fig. 16 muestra los niveles de actividad en una de esas redes después que ha sido entrenada. La red aprende ambos árboles familiares, esencialmente mediante la generalización de uno a otro, después de 1.500 presentaciones de los diversos tripletes. Una vez más, los campos receptivos de la unidad oculta se adaptan a la tarea. La fig. 17 muestra los campos de dos unidades

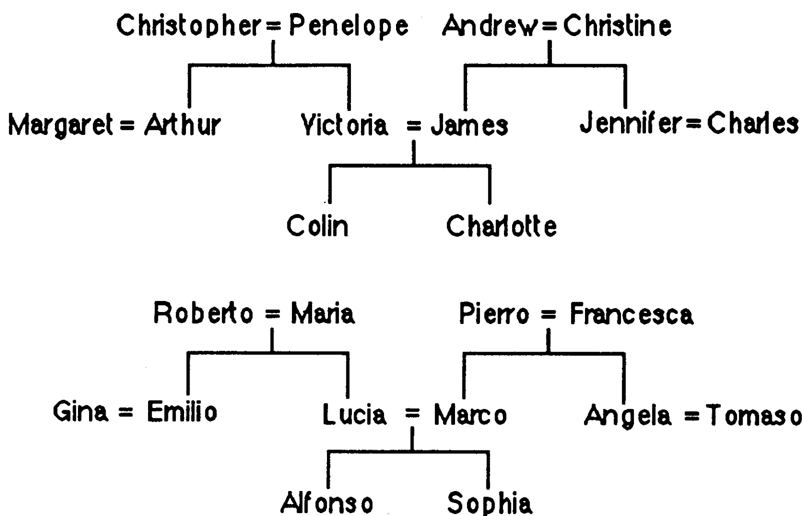


Figura 15. Dos árboles familiares isomorfos. La información puede expresarse como un conjunto de tripletes de la forma <persona 1> <relación> <persona 2>, donde las relaciones posibles son {padre, madre, esposo, esposa, hijo, hija, etc}. Se puede decir que una red de capas «conoce» esos tripletes si puede producir el tercer término de cada triplete cuando se le dan los otros dos. (Reproducido de David E. Rumelhart, Geoffrey E. Hinton y Robert J. Williams, «Learning Representations by Back-Propagating Errors», *Nature* 323 [1986]:533.)

en la primera capa oculta. La unidad 5 codifica la distinción entre inglés e italiano, mientras la unidad 6 codifica la rama de origen de la familia 1. Con esos campos receptivos en la unidad oculta, la red es capaz de generalizar correctamente cuando se le presentan tripletes nuevos.

Campos receptivos y neurobiología

Los ejemplos descritos muestran claramente que las unidades ocultas aprenden sobre los patrones de estímulo que se les presentan, sujeto a una retroalimentación orientada por la respuesta correcta de la red a tales patrones. En el aprendizaje sobre patrones de estímulo, las unidades ocultas desarrollan campos receptivos especializados (determinados por sus pesos de entrada) de una forma altamente cooperativa (cada campo está afectado

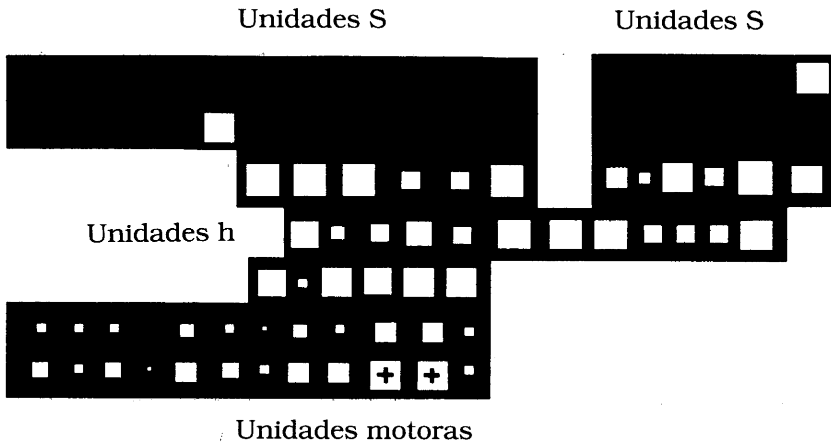


Figura 16. Niveles de actividad en una red de cinco capas después que ha sido entrenada. La capa superior tiene 24 unidades *S* en la izquierda para representar <persona 1> y 12 unidades *S* a la derecha para representar <relaciones>. Los cuadrados blancos muestran los niveles de actividad de las diversas unidades. Las unidades *S* activadas corresponden a <Colin> <tiene tía>. Cada uno de los dos grupos de entrada está conectado totalmente a su propio grupo de 6 unidades *h* en la segunda capa. Estos grupos aprenden a codificar personas y relaciones como patrones de actividad distribuidos. La segunda capa está totalmente conectada a la capa central de 12 unidades *h*, y éstas lo están a la siguiente cada capa de 6 unidades *h*. La actividad en este nivel debe activar las unidades motoras correctas, cada una de las cuales representa una <persona 2> particular. En este caso, hay dos respuestas correctas (marcadas con +), porque Colin tiene dos tías. (Reproducido de Rumelhart et al. [véase fig. 15].)

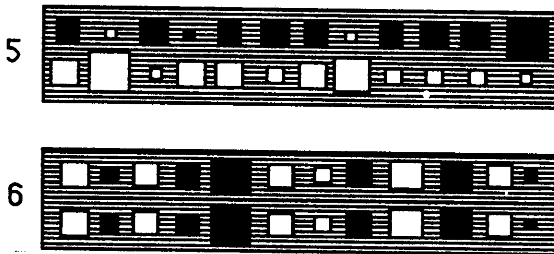


Figura 17. Campos receptivos de dos unidades ocultas que codifican información de árboles familiares (Reproducido de Rumerhart et al. [Véase fig. 15]).

por todos los demás campos de la red), y forman una representación distribuida de la clase de estímulo. Este resultado tiene importantes implicaciones para la neurobiología experimental,

porque aunque es sumamente improbable que en el cerebro se emplee algo parecido al procedimiento de propagación hacia atrás, el producto final (campos receptivos adaptados a una tarea dada) puede suministrar indicios sobre cómo podrían trabajar las redes neuronales reales ejecutando tareas similares. Estos indicios poseen consecuencias sumamente prácticas, dado que una de las pocas formas en que los neurobiólogos pueden observar directamente la operación del cerebro es determinando qué patrones de estímulo activan la neurona más próxima a un microelectrodo implantado en el cerebro.

Una primera aplicación de la propagación hacia atrás a lo largo de estas líneas ha sido reportada por Richard Andersen y David Zipser en conexión con los campos receptivos de neuronas en el córtex parietal posterior de un mono.[67] En esas neuronas, la información sobre la posición del ojo en relación con la cabeza y a la posición de la imagen en la retina se combina para localizar objetos con respecto a la posición de la cabeza en el espacio. Las neuronas en esta área poseen campos receptivos que se solapan en gran medida, y sus respuestas se coordinan de tal manera que se excitan a índices proporcionales a la diferencia entre la posición del ojo preferida y la posición real. Andersen y Zipser modelizaron esta situación con un perceptrón de tres capas con propagación hacia atrás entrenado para aprender posiciones espaciales. Después de menos de 1.000 intentos, las respuestas de la unidad motora codificaron adecuadamente las ubicaciones de los objetos en coordenadas centradas conforme la cabeza, dadas las coordenadas de la imagen del ojo y la retina. La red aprendió a asociar las posiciones de la imagen del ojo y la retina, y por supuesto los campos receptivos de la unidad oculta reflejaron esto. Lo que es más importante, estos campos receptivos se parecen a los del córtex parietal del mono. Es evidente que los perceptrones de capas múltiples pueden usarse en muchas situaciones similares para predecir las propiedades de los campos receptivos.

Extensión del procedimiento de propagación hacia atrás

El procedimiento de propagación hacia atrás descrito representa un importante avance en la teoría de los perceptrones y adalines, y más generalmente en la teoría y la práctica del

aprendizaje supervisado. Hay, sin embargo, unos pocos problemas con ese procedimiento. Primero, aunque es mucho más rápido que el procedimiento de Monte Carlo utilizado en la máquina de Boltzmann, la propagación hacia atrás sigue siendo un poco lenta. Se está dedicando mucho trabajo para encontrar procedimientos más rápidos y eficientes.[68] Otro problema concierne a la performance sobre problemas muy grandes. No siempre se sigue que los procedimientos que trabajan bien sobre problemas pequeños se comportarán también así sobre versiones más grandes del mismo problema. Esto se conoce como el problema del escalamiento. Han aparecido recientemente dos estrategias que enfrentan este problema.

En la primera, Dana H. Ballard señala que a veces los perceptrones de múltiples capas que utilizan propagación hacia atrás quedan «atrapados» en un patrón de actividad que no resuelve la tarea dada, particularmente en redes con más de una capa de unidades ocultas.[69] La solución de Ballard a esta situación es doble. Por un lado, construir módulos de aprendizaje asociativos en los cuales la salida de las unidades ocultas realimentan a las unidades S. El aprendizaje en una de esas redes no es supervisado, y las unidades ocultas forman una representación de las correlaciones existentes en y entre los patrones de las unidades S, lo mismo que en la máquina de Boltzmann. Por el otro, usar esas redes autoasociativas como módulos en una red jerárquicamente organizada. Estos módulos pueden acoplarse en formas que permitan resolver problemas cada vez mayores, resolviendo el problema del escalamiento. En un trabajo relacionado, Hinton y sus colegas han desarrollado un procedimiento similar, llamado recirculación, que minimiza el índice de cambio en la actividad de la red y que afirma ser neurobiológicamente más plausible que la propagación hacia atrás.[70]

Erich Mjolsness y David H. Sharp han introducido recientemente otro procedimiento que ha demostrado escalar muy bien en ciertos problemas y que se basa en el hecho de que la conectividad de una amplia clase de redes puede especificarse recursivamente, es decir, mediante la aplicación repetida de unas pocas reglas simples.[71] Esas redes se pueden modificar mediante cambios de reglas, más que mediante cambios de pesos. Además, el procedimiento está diseñado para penalizar redes con un gran número de contactos, de modo de aumentar la probabilidad de que esas redes

generalizarán a conjuntos de entrada más grandes. Esta estrategia se relaciona estrechamente con la de los llamados algoritmos genéticos, en los que las reglas para generar redes adaptativas están ellas mismas sujetas a adaptación.[72] El punto es que las redes genéticas escalan mejor que las redes especificadas directamente por sus pesos.

Redes amo-esclavo

Hoy en día existen muchas otras variantes de las redes de Hopfield, las máquinas de Boltzmann y los perceptrones de capas múltiples.[73] Un ejemplo muy interesante es el de la red «amo-esclavo» presentada recientemente por Alan S. Lapedes y Robert M. Farber.[74] Como ya hemos señalado, las redes de Hopfield (y las máquinas de Boltzmann) no son realistas como modelos de las redes neuronales, dado que todas sus conexiones son simétricas. Sin embargo, la habilidad de esas redes para encontrar configuraciones globalmente estables es valiosa. A la inversa, los perceptrones de capas múltiples no necesitan estar conectados simétricamente, pero su operación es sincrónica: no poseen ninguna dinámica intrínseca, aparte de la que proporciona un reloj externo. La red amo-esclavo introducida por Lapedes y Farber posee lo mejor de los dos mundos. El esclavo es una red de neuronas asincrónica y asimétrica, cuyos pesos son controlados y modificados por una red amo de Hopfield. Los diferentes estados estables de la red amo se pueden usar para controlar los pesos de la red esclava de manera tal de codificar cualquier comportamiento dinámico deseado. Las redes amo-esclavo son, de hecho, generalizaciones dinámicas de perceptrones de múltiples niveles, y se pueden usar para codificar representaciones, no sólo como los patrones y correlaciones estáticos de las unidades S, sino también como los patrones de unidades S que cambian en el tiempo.

Redes neuronales para predicción y simulación

Pero si las redes neuronales pueden aprender a representar correlaciones variables en el tiempo, se las puede utilizar como predictores y simuladores de una variedad de procesos dinámi-

cos. Lapedes y Farber han extendido recientemente su investigación de la propagación hacia atrás generalizada en esta dirección y han demostrado la eficacia de las redes neuronales para esos propósitos.[75] Con este desarrollo, la teoría de los perceptrones y los adalines ha renovado el contacto con la cibernética, la rama de la ingeniería que trata de la predicción, filtrado y simulación de procesos dinámicos.[76]

A este respecto, es de considerable interés comparar la estructura de un perceptrón de capas múltiples con la del filtro de aprendizaje inventado en 1954 por Dennis Gabor,[77] uno de los tempranos pioneros de la teoría de la comunicación y la cibernética y el inventor de la holografía. Este filtro actúa de la forma siguiente. Se almacena en cinta magnética una muestra grande de un mensaje ruidoso y se lo envía periódicamente a través del filtro. Se introducen en un comparador la salida del filtro y una copia avanzada o retardada del mensaje, lo que genera una diferencia entre las dos señales. Esta diferencia se usa entonces para ajustar el filtro de manera de minimizar la diferencia. De este modo la salida del filtro se termina pareciendo a la muestra del mensaje. Gabor demostró que una máquina semejante se podía entrenar para predecir y filtrar mensajes de diversas clases y también para reconocer patrones. El filtro ajustable es claramente análogo a una red de unidades *S*, *h* y motoras. (¡Parece evidente que Gabor resolvió el equivalente del problema de la asignación de crédito en 1954!) Lapedes y Farber reconocieron la analogía y demostraron que los perceptrones de capas múltiples con propagación hacia atrás generalizada son simuladores mucho más poderosos y flexibles que el filtro de aprendizaje de Gabor.

Las redes neuronales y los estudios de inteligencia artificial

En esta sección, que se basa en los materiales ya presentados en el ensayo, discutimos la relación entre las redes neuronales y la investigación en inteligencia artificial.

Consideramos la IA como el intento para corporizar en maquinaria computacional un repertorio de conducta inteligente comparable al comportamiento humano en contextos similares. Hasta hace poco este intento se llevó a cabo casi enteramente dentro del

paradigma estándar de la IA: primero especificar el contexto, después describir la lógica de la conducta deseada, y luego tratar de alcanzarla utilizando diversas heurísticas (es decir, métodos de búsqueda basados en un fondo de conocimiento previo suministrado por el diseñador).[78] Es evidente que un sistema así sólo tendrá éxito si el diseñador ha analizado la clase de problemas a resolver y es capaz de representar esta clase y las heurísticas de resolución del problema en un lenguaje de programación adecuado. No está claro si esta estrategia puede tener éxito en situaciones que tienen que reanalizarse porque el contexto ha cambiado. Parte del problema reside en la necesidad de descripciones lógicas sensibles al contexto. Es aquí donde las redes neuronales (más específicamente, perceptrones de capas múltiples) devienen relevantes. El diseñador de un perceptrón de capas múltiples no precisa una descripción lógica rigurosa; sólo necesita una comprensión informal de las complejidades de la conducta deseada, suficiente para construir la arquitectura global de una red neuronal apropiada. La propagación hacia atrás se encarga del resto de los detalles. De este modo, dado un tosco cableado general, los pesos de las conexiones locales se pueden cablear en fino mediante el entrenamiento. Las redes neuronales resultantes corporizan una descripción implícita de la conducta deseada, más que las afirmaciones lógicas declarativas explícitas que controlan un sistema de IA. A este respecto, es oportuno recordar una afirmación de John von Neumann, concerniente a la pregunta de si toda conducta se puede expresar completa y no ambiguamente en palabras y por lo tanto en símbolos lógicos. Von Neumann preveía que el problema de la visión era inmensamente complicado:

> No es cierto del todo que... [un objeto visual] podría no constituir la descripción más simple de sí mismo... [Más aún,] no es... del todo improbable que sea fútil buscar un concepto lógico preciso, es decir, una descripción verbal precisa [digamos] de «analogía visual». Es posible que el patrón de conexión del cerebro visual mismo sea la expresión lógica o la definición más simple de este principio.[79]

Por supuesto, los perceptrones de capas múltiples no han sido utilizados aún para la resolución de problemas del mismo modo que lo ha sido la IA, y pudiera suceder que a medida que se ataquen problemas más difíciles, las complejidades del cableado

global demostraran ser demasiado formidables. A este respecto, los intentos de Ballard, Hinton, Mjolsness y Sharp para construir redes jerárquicamente organizadas cableadas por soft pueden resultar útiles. En todo caso, parece razonable esperar que los perceptrones de capas múltiples sean capaces de corporizar descripciones implícitas de toda conducta complicada, siempre que alguna comprensión previa de la estructura lógica de esa conducta se codifique de algún modo en la arquitectura global.

Memoria

También está el problema de la memoria. Uno de los factores que limitan el desarrollo de los sistemas de IA es el costo de las memorias amplias y rápidas.[80] Los ACAMs en forma de redes de Taylor-Steinbuch, o en su forma moderna como redes de Marr,[81] o más probablemente en forma de redes de Hopfield mejoradas, jugarán probablemente un papel esencial en proporcionar esas memorias, particularmente a la luz del desarrollo de circuitería integrada en gran escala (VLSI), la cual se puede microcablear para toda una variedad de tareas.[82] Esos almacenamientos de memoria, en combinación con las redes susceptibles de ser entrenadas que hemos descrito, bien pueden proporcionar un sustrato adecuado para encarnar en máquinas conducta verdaderamente inteligente.

Perspectivas futuras

¿Podemos esperar, por lo tanto, ver robots autónomos e inteligentes, con cerebros de silicio, construidos a partir de perceptrones de múltiples capas y ACAMs jerárquicamente organizados y mejorados, en un futuro no muy distante? Es nuestra creencia, basada en las investigaciones que hemos descrito en este ensayo, que todavía falta recorrer un largo camino antes que pueda producirse cualquier clase de robot verdaderamente inteligente. Está claro que todo el progreso hasta la fecha en el cableado de redes para ejecutar tareas inteligentes descansa en análisis previos por parte del diseñador de la tarea dependiente del contexto que debe desarrollarse. Es el diseñador quien aporta la

intención y el significado. La IA descansa en la idea expresada en forma muy sucinta por Kenneth J. W. Craik en 1943, de que «el pensamiento reproduce la realidad mediante el simbolismo».[83] Pero ya hemos señalado que para muchos e importantes procesos de pensamiento pueden no existir descripciones simbólicas explícitas, de modo que ¿pueden las redes neuronales entrenables «describir» esos procesos? La respuesta parece ser que sí, pero el entrenador necesita saber muchísimo sobre la estructura de esos procesos.

¿Qué hay de las redes de Hopfield y de las máquinas de Boltzmann, que aprenden sin entrenadores? Dada la restricción de los pesos simétricos, éstas parecen ser más promisorias como modelos de la forma en que trabaja el cerebro; muchos investigadores han desarrollado redes de Hopfield con pesos sinápticos modificables para este propósito. Christoph von der Malsburg, por ejemplo, ha utilizado esas redes para modelizar el pensamiento asociativo, en el que se producen nuevas combinaciones de representaciones por medio de rápidas modificaciones sinápticas.[84] Recientemente se ha llamado sinapsis de Malsburg a esas sinapsis.[85] No está claro, sin embargo, que las redes de Malsburg resuelvan el problema de la asignación de crédito. Las máquinas de Boltzmann ciertamente lo hacen, pero sólo si hay una red externa oculta que indique a las unidades ocultas que la red está siendo estimulada. (Como hemos dicho, el procedimiento de aprendizaje de la máquina de Boltzmann requiere que las unidades ocultas conozcan la diferencia entre la actividad libre y la estimulación.) Este es un modelo primitivo de atención. Que se necesita ese mecanismo para controlar el aprendizaje y el pensamiento no es, por cierto, una novedad. Pensando en el cerebro, Francis Crick ha sugerido que ese mecanismo existe no en el neocórtex cerebral sino en el tálamo, la parte del cerebro que se encuentra entre el neocórtex y la raíz del cerebro y que contiene todas las redes que controlan las funciones corporales. La necesidad de un mecanismo semejante en las redes de Hopfield ya ha sido apreciada por Mjolsness como una forma de serializar el proceso de búsqueda que encuentra configuraciones estables.[86] ¿Pueden estas redes reemplazar a los perceptrones de capas múltiples con propagación hacia atrás? Según creemos, es probable que la respuesta sea positiva, pero una vez más, sólo si en esa red ya hay una gran cantidad de cableado previo.

Es ese cableado duro el que corporiza el conocimiento previo y,

en cierto sentido, la intención del diseñador. En el cerebro humano el cableado duro es el producto final de mil millones de años de adaptaciones evolutivas ante ambientes cambiantes, expresadas mediante la acción de productos genéticos durante el curso del desarrollo del cerebro. En cierto sentido, la evolución ha actuado no como un entrenador para el cableado en suave de redes neuronales, sino como un crítico para cablearlas en duro: si trabaja, sobrevive, de modo que el cableado en suave ulterior es efectivo. ¿Debemos esperar ser capaces de concentrar mil millones de años de evolución actuando sobre el protoplasma en unas pocas décadas de investigación en redes neuronales e IA sobre chips de silicio VLSI? Hasta que comprendamos la forma en que toman cuerpo las ideas y las intenciones en el cerebro humano, el progreso veloz parece improbable. Por otro lado, los desarrollos en la teoría y en la práctica de los perceptrones de capas múltiples que hemos descrito permiten la investigación experimental del cableado duro mismo. Predecimos que la estrategia de arriba hacia abajo de la IA convencional y la estrategia de abajo hacia arriba del neoconexionismo pueden unirse eventualmente para producir un progreso real en lo que McCulloch una vez llamó epistemología experimental: el estudio de la forma en que el conocimiento se encarna en el cerebro y podría encarnarse en las máquinas.[87]

Notas

Jack Cowan agradece la hospitalidad del Centro para la Biología Matemática, Instituto de Matemáticas, Universidad de Oxford; el Instituto para la Física Teórica de la Universidad de California en Santa Barbara, y el Departamento de Matemáticas de la Universidad del Sur de California, así como la ayuda financiera proporcionada en parte por el Laboratorio Nacional de Los Alamos, el Consejo de Investigación en Ciencia e Ingeniería de Gran Bretaña (Beca Nº GR/D/13573), la Fundación Nacional de la Ciencia (Beca Nº PHY82-17853, suplementada por fondos de la Academia Nacional de Aeronáutica y del Espacio), la Fundación para el Desarrollo de Sistemas (Beca Nº SDF 55) y el Programa en Ciencia Neuronal, Informacional y de la Conducta (NIBS) de la Universidad del Sur de California. David Sharp agradece a la Universidad de Chicago y al Instituto para la Física Teórica, a la Universidad de California en Santa Barbara, por su hospitalidad, y al Departamento de Energía de los EE.UU. por su ayuda financiera.

[1] Warren S. McCulloch y Walter H. Pitts, «A Logical Calculus of the Ideas Immanent in Nervous Activity», *Bulletin of Mathematical Biophysics* 5 (1943):115.

[2] Donald M. Mackay, «On Comparing the Brain with Machines», *American Scientist* 42 (1954):2.

[3] Alan M. Turing, «On Computable Numbers with an Application to the Entscheidungsproblem», en *Proceedings of the London Mathematical Society* 42 (1936):230; 43 (1937):544.

[4] John von Neumann, «Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components», en *Automata Studies*, ed. de C.E. Shannon y John McCarthy (Princeton, N.J.: Princeton University Press, 1956).

[5] Shmuel Winograd y Jack D. Cowan, *Reliable Computation in the Presence of Noise* (Cambridge, MIT Press, 1963).

[6] J. Taylor, *Selected Writings of John Hughlings Jackson* (Londres, Hodder y Stoughton, 1932).

[7] Karl S. Lashley, «Persistent Problems in the Evolution of Mind», *Quarterly Review of Biology* 24 (1) (1942):28.

[8] Karl S. Lashley, «In Search of the Engram», *Symposium of the Society for Experimental Biology* 4 (1950):454.

[9] Horace B. Barlow, «Single Units and Sensation: A Neuron Doctrine for Perceptual Psychology?», *Perception* 1 (1972):371.

[10] Donald O. Hebb, *The Organization of Behavior* (Nueva York, John Wiley, 1949).

[11] Albert M. Uttley, «The Classification of Signals in the Nervous System», *EEG Clinical Neurophysiology* 6 (1954):479.

[12] Walter H. Pitts y Warren S. McCulloch, «How We Know Universals: The Perception of Auditory and Visual Forms», *Bulletin of Mathematical Biophysics* 9 (1947):127.

[13] O. Lippold, *The Origin of the Alpha Rhythm* (Edimburgo y Londres: Churchill Livingstone, 1953).

[14] Donald A. Mackay, «Some Experiments on the Perception of Patterns Modulated at the Alpha Frequency», *EEG Clinical Neurophysiology* 5 (1953):559.

[15] Frank Rosenblatt, «The Perceptron, a Probabilistic Model for Information Storage and Organization in the Brain», *Psychological Review* 62 (1958):386.

[16] A. Novikoff, «On convergence Proofs for Perceptrons», en *Symposium on the Mathematical Theory of Automata*, ed. J. Fox (Nueva York, Polytechnic Press, 1963):615.

[17] Bernard Widrow y M.E. Hoff, «Adaptive Switching Circuits», *WESCON Convention Record* 4 (1960):96.

[18] Marvin Minsky y Seymour Papert, *Perceptrons: An Introduction to Computational Geometry* (Cambridge, MIT Press, 1969).

[19] Rosenblatt, «The Perceptron».

[20] Wilfrid K. Taylor, «Electrical Simulation of Some Nervous System Functional Activities», en *Information Theory*, ed. E. C. Cherry (Londres, Butterworths, 1956), 3.

[21] G.S. Brindley, «The Classification of Modifiable Synapses and Their Use in Models for Conditioning», en *Proceedings of The Royal Society of London* B (1967):168, 361.

[22] R.A. Rescorla y A. R. Wagner, «A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement», vol. 2 de *Classical Conditioning*, ed. A.H. Black y W.F. Prokasy (Nueva York: Appleton-Century-Crofts, 1972).

[23] Wilfrid K. Taylor, «Cortico-Thalamic Organization and Memory», en *Proceedings of the Royal Society of London* B (1964):159, 466.

[24] Karl Steinbuch, «Die Lernmatrix,» *Kybernetik* 1 (1) (1961):36.

[25] James A. Anderson, «A Memory Storage Model Utilizing Spatial Correlation Functions,», *Kybernetik* 5 (1968):113.

[26] David J. Willshaw, O. Peter Buneman y H. Christopher Longuet-Higgins, «Non-holographic Associative Memory,» *Nature* 222 (1969):960.

[27] David Marr, «A Theory of Cerebellar Cortex», *Journal of Physiology of London* 202 (1969):437; Marr, «Simple Memory: A Theory for Archicortex», *Philosophical Transactions of the Royal Society of London* B 262 (841) (1971):23.

[28] Teuvo Kohonen, *Associative Memory: A System-Theoretical Approach* (Berlín, Springer-Verlag, 1977).

[29] J.C. Eccles, M. Ito y J. Szentágothai, *The Cerebellum as a Neuronal Machine* (Nueva York: Springer-Verlag, 1967); R. Llinás, ed., *Neurobiology of Cerebellar Evolution and Development* (Chicago: American Medical Association, 1969).

[30] Marr, «A Theory of Cerebellar Cortex».

[31] S. Blomfield y David Marr, «How the Cerebellum May Be Used», *Nature* 227 (1970):1224.

[32] James S. Albus, «A Theory of Cerebellar Function», *Mathematical Bioscience* 10 (1971):25.

[33] M. Ito. *The Cerebellum and Neural Control* (Nueva York: Raven, 1984); P.L. Strick, *Science* 229 (1985):547.

[34] Marr, «Simple Memory: A Theory of Archicortex».

[35] J. O'Keefe y L. Nadel, *The Hippocampus as a Cognitive Map* (Oxford: Clarendon Press, 1978).

[36] David Marr, «A Theory for Cerebral Neocortex», *Proceedings of the Royal Society of London* B (176) (1970):161.

[37] T.V.P. Bliss y T. Lomo, «Long-Lasting Potentiation of Synaptic Transmission in the Dentate Area of the Unaesthetized Rabbit Following Stimulation of the Perforant Path», *Journal of Physiology of London* 232 (1973):357.

[38] David J. Willshaw y Christoph von der Malsburg, «How Patterned Neural Connections Can Be Set Up by Self-Organization», en *Proceedings of the Royal Society of London* B (194) (1976):431; D.J. Willshaw y Ch. von der Malsburg, «A Marker Induction Mechanism for the Establishment of Ordered Neural Mappings: Its Application to the Retinotectal Problem», *Philosophical Transactions of the Royal Society of London* B (287) (1979):203; S.E. Fraser y R.K. Hunt, «Retinotectal Specificity: Models and Experiments in Search of a Mapping Function», *Annual Review of Neuroscience* 3 (1980):319; V.A. Whitelaw y Jack D. Cowan, «Specificity and Plasticity of Retinotectal Connections: A Computational Model», *Journal of Neuroscience* 1 (12) (1981):1369.

[39] E.M. Hart et al., «Brain Functions and Neural Dynamics», *Journal of Theoretical Biology* 26 (1970):93; H.R. Wilson y Jack D. Cowan, «Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons», *Biophysics Journal* 12 (1972):1; H.R. Wilson y Jack D. Cowan, «A Mathematical Theory of the Functional Dynamics of Cortical and Thalamic Nervous Tissue», *Kybernetik* 13 (1973):55; S. Grossberg, «Contour Enhancement, Short-term Memory, and Constancies in Reverberating Neural Networks», *Studies in Applied Mathematics* 52 (3) (1973):213; G.B. Ermentrout y Jack D. Cowan, «Temporal Oscillations in Neural Networks», *Journal of Mathematical Biology* 7 (1979):265; «A Mathematical Theory of Visual Hallucination Patterns», *Biological Cybernetics* 34 (1979):137.

[40] Santiago Ramón y Cajal, *Histology du Système Nerveux* (reimpresión: Madrid: Consejo Superior de Investigaciones Científicas, 1972).

[41] John J. Hopfield, «Neural Networks and Physical Systems with Emergent Collective Computational Abilities», *Proceedings of the National Academy of Sciences* 79 (1982):2554; Hopfield, «Neurons with Graded Response Have Collective Computational Properties Like Those of Two-State Neurons», *PNAS* 81 (1984):3088.

[42] Si *x* e *y* son dos neuronas, entonces sus conexiones son simétricas si el peso de la sinapsis de *x* a *y* iguala al peso de la sinapsis de *x* a *y*.

[43] David Sherrington y Scott Kirkpatrick, «Spin Glasses», *Physics Review Letters* 35 (1975):1972; S.F. Edwards y P.W. Anderson, «Theory of Spin-Glasses: I», *Journal of Physics F: Metal Physics* 6 (10) (1976):1927.

[44] Brian G. Cragg y H. Nevill V. Temperley, «The Organisation of Neurones: A Cooperative Analogy», *EEG Clinical Neurophysiology* 6 (85) (1954):37.

[45] Brian G. Cragg y H. Nevill V. Temperley, «Memory: The Analogy with Ferromagnetic Hysteresis», *Brain* 78 (2) (1955):304.

[46] William A. Little, «The Existence of Persistent States in the Brain», *Mathematical Bioscience* 19 (1974):101.

[47] Sherrington y Kirkpatrick, «Spin Glasses».

[48] La regla de Hebb se puede exponer como sigue. Supongamos que las neuronas *x* e *y* están conectadas. Hagamos que $X=\pm 1$ sea el estado de la neurona *x*, e $Y=\pm 1$ el estado correspondiente de la neurona *y*. Luego el peso sináptico del contacto de *x* a *y* es proporcional al valor promedio del producto XY. Este es un ejemplo de una regla local. El peso está determinado sólo por las actividades correlacionadas de las neuronas *x* e *y*.

[49] W. Ross Ashby, «The Stability of a Randomly Assembled Nerve-Network», *EEG Clinical Neurophysiology* 2 (1950):471.

[50] S.-I. Amari, «Characteristics of Random Nets of Analog Neuron-like Elements», *Institute of Electronic and Electrical Engineers Transactions on Systems, Man, and Cybernetics*, SMC-2 (5) (1972):643; R. L. Beurle, «Properties of a Mass of Cells Capable of Regenerating Pulses», *Philosophical Transactions of the Royal Society of London* B (240) (669) (1956):55; Jack D. Cowan, «The Problem of Organismic Reliability», *Progressive Brain Research* 17 (1965):9; B.G. Cragg y H.N.V. Temperley, «The Organisation of Neurones»; S. Grossberg, «Contour Enhancement, Short-term Memory, and Constancies in Reverberating Neural Networks», *Studies in Applied Mathematics* 52 (3) (1973):213; W.A. Little y G.L. Shaw, «A Statistical Theory of Short and Long-Term Memory», *Behavioral Biology* 14 (1975):115; H.R. Wilson y Jack D. Cowan, «Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons», *Biophysics Journal* 12 (1972):1; H.R. Wilson y Jack D. Cowan, «A Mathematical Theory of the Functional Dynamics of Cortical and Thalamic Nervous Tissue», *Kybernetik* 13 (1973):55.

[51] E. L. Lawler et al., editores, *The Travelling Salesman Problem* (Nueva York: John Wiley, 1985).

[52] S. Kirkpatrick, C.D. Gelatt (h) y M. P. Vecchi, «Optimization by Simulated Annealing», *Science* 229 (4598) (1983):671.

[53] John J. Hopfield y David W. Tank, «'Neural' Computation and Constraint Satisfaction Problems and the Traveling Salesman», *Biological Cybernetics* 55 (1985):141.

[54] S. Lin y B.W. Kernighan, «An Algorithm for the TSP Problem», *Operations Research* 21 (1973):498.

[55] Richard Durbin y David J. Willshaw, «An Analogue Approach to the Travelling Salesman Problem Using an Elastic Net Method», *Nature* 326 (1987):689.

[56] Nicholas Metropolis et al., «Equations of State Calculations by Fast Computing Machines», *Journal of Chemical Physics* 21 (1953):1087.

[57] Geoffrey E. Hinton y Terrence J. Sejnowski, «Optimal Perceptual Inference», en *Proceedings of the Institute of Electronic and Electrical Engineers Computer Society on the Conference on Computer Vision and Pattern Recognition* (Washington, D.C.: IEEE, 1983); D. H. Ackley, G. E. Hinton y T. J. Sejnowski, «A Learning Algorithm for Boltzmann Machines», *Cognitive Science* 9 (1985):147.

[58] Dana H. Ballard, Geoffrey E. Hinton y Terrence J. Sejnowski, «Parallel Visual Computation», *Nature* 306 (5938) (1983):21.

[59] F.A. Hayek, *The Sensory Order* (Chicago: University of Chicago Press, 1952).

[60] Geoffrey E. Hinton, *Distributed Representations*, Technical Report CMU-CS-84-157 (Pittsburgh: Carnegie-Mellon University, Computer Science, 1984).

[61] David E. Rumelhart, Geoffrey E. Hinton y Robert J. Williams, «Learning Internal Representations by Error Propagation», *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, *Foundations*, ed. David E. Rumelhart y James L. McClelland (Cambridge: MIT Press, 1986); Y. Le Cun, «A Learning Scheme for Asymmetric Threshold Networks», en *ProceedingsCognitiva* 85 (1985):599; D.B. Parker, «A Comparison of Algorithms for Neuron-Like Cells», en *Proceedings of the AIP Conference* 151, *Neural Networks for Computing*, ed. J.S. Denker, AIP.

[62] L. Uhr, ed., *Pattern Recognition* (Nueva York: John Wiley, 1966).

[63] D. H. Ballard, «Cortical Connections and Parallel Processing», *Behavioral and Brain Sciences* 9 (1) (1986):67.

[64] Terrence J. Sejnowski y Charles R. Rosenberg, NETtalk: *A Parallel Network that Learns to Read Aloud*, Technical Reports JHU/EECS-86/01 (Baltimore, Md.: Johns Hopkins University, Electrical Engineering and Computer Science, 1986).

[65] Rosenblatt, «The Perceptron».

[66] David E. Rumelhart, Geoffrey E. Hinton y Robert J. Williams, «Learning Representations by Back-Propagating Errors», *Nature* (323) (1986):533.

[67] Richard Andersen y David Zipser, «A Neural Net Model of Posterior Parietal Cortex», en *Neurobiology of Neocortex* (Berlín: Dahlem Conferenz, 1987): 50; J. Altman, "A Quiet Revolution in Thinking", *Nature* 328 (1987): 572.

[68] W. S. Stornetta y B. A. Huberman, «An Improved Three-Layer Back-Propagation Algorithm», preimpreso de la Xerox Corporation, Palo Alto, 1987.

[69] Dana H. Ballard, «Modular Learning in Neural Networks», preimpreso de University of Rochester Department of Computer Science.

[70] G. North (1987), «A Celebration of Connectionism», *Nature* 328 (1987):107.

[71] Erich Mjolsness y David H. Sharp, «A Preliminary Analysis of Recursively Generated Networks», en *Proceedings of the AIP Conference on Neural Networks for Computing*, ed. J.S. Denker (1986): 151; E. Mjolsness, David H. Sharp y B.K. Alpert, «Recursively Generated Networks», YALEU/DCS/RR-549, preimpresión de la Universidad de Yale, 1987.

[72] J. H. Holland, *Adaptation in Natural and Artificial Systems* (Ann Arbor: University of Michigan Press, 1975).

[73] R. P. Lippmann, «An Introduction to Computing with Neural Nets», *IEEE ASSP Magazine* 4 (2) (1987):4.

[74] Alan Lapedes y Robert Farber, «A Self-Optimizing, Nonsymmetrical Neural Net for Content Addressable Memory and Pattern Recognition», *Physica D* 22 (1986):247.

[75] Alan Lapedes y Robert Farber,, «Nonlinear Signal Processing Using Neural

Networks: Prediction and System Modeling», LA-UR-87-2662, preimpreso de Los Alamos National Laboratory.

[76] Norbert Wiener, *Cybernetics, or Control and Communication in the Animal and the Machine* (Nueva York: John Wiley, 1948).

[77] Dennis Gabor, «Communication Theory and Cybernetics», en *IRE Transactions* CT-1 (4) (1954):19.

[78] David L. Waltz, «The Prospects for Building Truly Intelligent Machines», *Daedalus* (verano de 1988): 191-212.

[79] John von Neumann, «The General and Logical Theory of Automata», en *Cerebral Mechanisms in Behavior, the Hixon Symposium*, ed. L. A. Jeffress (Nueva York : John Wiley, 1951).

[80] Waltz, «The Prospects for Building Truly Intelligent Machines».

[81] P. Kanerva, *Self-Propagating Search: A Unified Theory of Memory* (Cambridge: MIT Press, en prensa).

[82] J. S. Denker, ed., *Proceedings of the AIP Conference on Neural Networks for Computing* (1986):151.

[83] Kenneth J. W. Craik, *The Nature of Explanation* (Cambridge: Cambridge University Press, 1943).

[84] Christoph von der Malsburg, «Nervous Structures with Dynamical Links», *Berichte Bunsenges Physikalische Chemie* 89 (1975):703.

[85] Francis Crick, «Function of the Thalamic Reticular Complex: The Searchlight Hypothesis», *PNAS* 81 (1984):4586.

[86] Eric Mjolsness, «Control of Attention in Neural Networks», YALEU/DCS/RR-545, preimpreso de la Universidad de Yale.

[87] Warren S. McCulloch, «A Historical Introduction to the Postulational Foundations of Experimental Epistemology», en F.S.C. Northrop y H.H. Livingston, editores, *Cross-Cultural Understanding: Epistemology in Anthropology* (Nueva York: Harper and Row, 1964).

6
El nuevo conexionismo: desarrollando relaciones entre la neurociencia y la inteligencia artificial

Jacob T. Schwartz

Parte de la confianza con que los investigadores en inteligencia artificial contemplan las perspectivas de su campo procede de los supuestos materialistas a los que suscriben. Uno es que la «mente» es simplemente un nombre para la actividad de procesamiento de información del cerebro. Otro es que el cerebro es una entidad física que actúa conforme a las leyes de la bioquímica y que no está influenciado por ningún «alma» irreducible o por ninguna otra entidad unitaria, puramente mental, imposible de analizar como secuencia causal de sucesos bioquímicos elementales. Esta perspectiva, aceptada ampliamente, junto con la masa de información en rápido crecimiento atinente a la fisiología del sistema nervioso, a la microanatomía y a la conducta de señalización, y junto con el impulso, basado en la tecnología actual, que se ha dado a la construcción de sistemas de computación analógicos que involucran miles de elementos actuando en paralelo, ha alimentado un nuevo acento entre los investigadores de IA que ha llegado a conocerse como «el nuevo conexionismo». Las premisas enfatizadas que caracterizan a esta nueva escuela de pensamiento rezan como sigue:

1. El cerebro no opera como una computadora serial de tipo convencional, sino en forma masivamente paralela. El funcionamiento paralelo de cientos de miles o millones de neuronas en los sutiles procesos de extracción de información del cerebro proporcio-

---

Jacob T. Schwartz. Profesor del Instituto Courant de Ciencias Matemáticas de la Universidad de Nueva York.

na velocidad. Los perceptos coherentes se forman en tiempos que exceden a los tiempos elementales de reacción de las neuronas singulares por un factor de poco más de diez. Especialmente en lo que concierne a los procesos perceptuales básicos, como la visión, esta observación excluye las formas de procesamiento de información iterativas que tendrían que barrer los datos que ingresan serialmente, o pasarlos a través de muchas etapas de procesamiento intermediarias. Dado que las operaciones de búsqueda simbólica seriales y extensivas de este tipo no parecen caracterizar el funcionamiento de los sentidos, el supuesto (típico de gran parte de las especulaciones de la ciencia cognitiva inspirada en la IA en el período 1960-1980) de que la búsqueda serial subyace a las diversas funciones cognitivas más elevadas deviene sospechoso.

2. Dentro del cerebro, el conocimiento no se almacena en ninguna forma que se parezca a un programa de computadora convencional, sino que se almacena estructuralmente, en forma de patrones distribuidos de pesos sinápticos excitatorios e inhibitorios, cuyas magnitudes relativas determinan el flujo de las respuestas neuronales que constituyen la percepción y el pensamiento.

Los investigadores de IA que desarrollan estos puntos de vista se han involucrado con la neurociencia esperando poder contribuir con vislumbres teoréticas que aporten significado a una masa de datos empíricos en rápido crecimiento, pero todavía desconcertante, reunida por los neurocientíficos experimentales (muchos de los cuales consideran la especulación teorética con algo más que un poco de desdén). Estos investigadores en IA esperan combinar indicios procedentes de la experimentación con la habilidad práctica que tienen los computadores científicos para analizar complejas funciones externas en patrones de acciones elementales. Aceptando el supuesto de que existe alguna forma general para las actividades computacionales características de esas acciones, esperan adivinar algo iluminador sobre la forma en que surgen los procesos perceptuales y cognitivos del cerebro. Es decir, los científicos computacionales esperan relacionarse con la neurociencia experimental en forma muy parecida a la que se relacionan la física teórica y la física experimental: contribuyendo a unificar percepciones teóricas y conjeturas basadas teoréticamente que puedan guiar la experimentación a lo largo de caminos fructíferos.

La aterradora complejidad del cerebro opone grandes obstáculos a una fácil materialización de este objetivo. Sugeriré la magnitud de los problemas que requieren resolverse mediante unas pocas estimaciones intimidantes y una breve revisión de algunos hechos básicos de la neurociencia. El cerebro humano consiste en aproximadamente 100 mil millones de neuronas, que posiblemente sean diez veces más. Las neuronas se comunican habitualmente transmitiendo paquetes eléctricos discretos (potenciales de acción) a una población de neuronas vecinas. Por lo que se sabe, la amplitud precisa y la forma de esos paquetes y el tiempo preciso para su llegada en un intervalo de dos milisegundos o algo así, son detalles físicos que el sistema nervioso no es capaz de explotar. De allí que se pueda modelizar cada paquete como un «bit» singular portador de información en la corriente de salida de una neurona y decir que una neurona produce información a un ritmo de aproximadamente cien bits por segundo. Esta forma de pensar lleva a un estimado de 10 billones de bits por segundo, tomando un factor de cien, para el «ancho de banda» interno del cerebro.

La actividad computacional de cada neurona involucra una gran cantidad de mecanismos, aún imperfectamente conocidos. Sin embargo, una masa considerable de evidencia experimental sustenta el siguiente cuadro general. Una neurona transmite información a las neuronas vecinas en uniones neuronales llamadas sinapsis. Una sola neurona puede tener tanto como diez mil entradas sinápticas, aunque en algunos casos muchas entradas menos, y en otros casos tanto como 100 mil entradas convergen en una sola neurona. El número total de sinapsis en el cerebro se puede estimar en 1.000 billones, aunque esta estimación, igual que las ofrecidas en los siguientes párrafos, es incierta en un factor de 100. Las señales de entrada transmitidas a una neurona (por lo general químicamente) a través de una sinapsis disparan una amplia variedad de reacciones. Una es la modulación de la conductividad iónica de la membrana de la neurona afectada, que ya sea eleva el voltaje de una porción de su interior (excita a la neurona) o disminuye su voltaje (la inhibe). Después de una atenuación en el espacio y en el tiempo, que se manifiesta en función de la química y la geometría de la neurona afectada y de sus sinapsis, la neurona combina el cambio de voltaje generado por esos efectos sinápticos. Si el voltaje combinado resultante (sumado, por ejemplo) excede el umbral de la reacción, la neurona

genera un paquete de salida u otra señal eléctrica. Esto se transmite luego a todas sus sinapsis de salida. Aunque muchos otros mecanismos juegan su papel, este tipo de efecto parece básico para muchas de las computaciones más rápidas ejecutadas por el cerebro.

Se sabe que otras formas de insumos sinápticos poseen efectos bioquímicos más lentos pero más perdurables que esos efectos iónicos, que probablemente sustenten el núcleo de la actividad de trasmutación de información cerebral. La estimulación de ciertas sinapsis puede, por ejemplo, disparar en una neurona actividades enzimáticas que modifican sus actividades biosintéticas, por ejemplo, aumentando o disminuyendo su susceptibilidad a los estímulos excitatorios o inhibitorios que actúan iónicamente. Dependiendo de los efectos químicos involucrados, esas modificaciones sinápticas de rápidas respuestas iónicas pueden durar tan poco como cincuenta milisegundos o tanto como varios segundos, minutos o días; incluso pueden ser permanentes. Otras reacciones enzimáticas sinápticamente desencadenadas pueden iniciar cambios bioquímicos en secuencia. Por ejemplo, la respuesta eléctrica de una neurona se puede aumentar por varias decenas de milisegundos pero luego se puede inhibir por un período mayor, lo que conduce a complejos patrones de alternancia entre excitación e inhibición. La variedad de conductas de neuronas singulares que puede engendrar el amplio espectro de acciones enzimáticas ha sido estudiada en animales simples como la *Aplysia*, de algunas de cuyas neuronas se sabe que poseen patrones enormemente individualizados de actividad continua, periódica o en ráfagas.

Aunque no es fácil resumir la amplia variedad de patrones de respuesta sináptica con unos pocos números que representen el poder de procesamiento de información y la capacidad de almacenamiento de una sola neurona, las estimaciones siguientes no son deshonestas. Un byte (ocho bits, alrededor de un carácter de impresión) puede ser suficiente para representar la fuerza a largo plazo de una sinapsis. Se deben tomar cuatro bytes adicionales para proporcionar una representación completa del estado bioquímico a corto plazo de ambos lados de una sinapsis y del estado del vacío sináptico correspondiente, determinado por su historia de estimulación hasta un momento dado. Estas gruesas estimaciones nos llevan a considerar que la memoria a largo plazo

disponible en el cerebro es de alrededor de 1.000 billones de bytes y que la cantidad de datos a corto plazo que se necesita para caracterizar el estado de cada una de sus sinapsis es aproximadamente la misma. La actividad lógica de cada neurona puede considerarse como un proceso que combina aproximadamente 10 mil bytes de entrada con unos 40 mil caracteres de estado de sinapsis a un ritmo de 100 veces por segundo. La cantidad de aritmética analógica requerida por esta estimación es (de nuevo, muy rudamente) de 10 millones de operaciones elementales por neurona por segundo, sugiriendo que el ritmo de computación que se necesita para emular todo el cerebro sobre una base de neurona por neurona puede ser tan alta como 1.000.000 billones de operaciones aritméticas por segundo. (Por supuesto, ritmos de computación varios órdenes de magnitud menores pueden ser suficientes para representar el contenido lógico de la actividad del cerebro, si es que llega a descubrirse cuál es).

Es interesante comparar estas estimaciones excesivamente gruesas con las cifras correspondientes de los sistemas de supercomputadora más grandes que probablemente se desarrollen la próxima década. Estos probablemente no excedan mucho la velocidad de 1 billón de operaciones aritméticas por segundo, lo que es alrededor de la millonésima parte del ritmo de computación que hemos estimado para el cerebro. Los discos magnéticos más grandes de la actualidad almacenan alrededor de 1.000 millones de bytes de información cada uno, lo que es groseramente la diez millonésima parte de la capacidad de almacenamiento que hemos adscripto al cerebro. Aun si suponemos rápidos avances en la tecnología de almacenamiento y sistemas equipados con centenares de discos, parece improbable que las supercomputadoras alcancen más del 1 por ciento de la capacidad de almacenamiento del cerebro en la próxima década. Claramente, los neurocientíficos confrontan un sistema cuyo trabajo es difícil de evaluar físicamente y cuyas operaciones son de una estremecedora complejidad.

Indicios sobre las funciones cerebrales

Uno de los indicios más salientes a partir del cual esperan trabajar los teóricos «conexionistas» es la observación de que los procesos mentales (especialmente los sensoriales) parecen ser de

una «profundidad» muy restringida, en el sentido de que no se requieren muchas reacciones neuronales elementales sucesivas para formar las reacciones del más alto nivel que el cerebro genera. Simplemente no hay tiempo de involucrar muchas reacciones sucesivas.

Esta es sólo una pista débil, sin embargo. Dado que las neuronas son harto más complejas que los conmutadores elementales que se usan para construir computadoras, una sola etapa de procesamiento neuronal se puede comparar a diez o más etapas de procesamiento electrónico por parte de los componentes elementales del tipo de los conmutadores. Por consiguiente, los eductos que puede generar el cerebro en un décimo de segundo se pueden comparar en complejidad con los eductos que requieren cien o más etapas de procesamiento en los conmutadores electrónicos. Más aún, se comprende tan poco la significación lógica de las interconexiones en el sistema nervioso (aun en los casos en que conocemos bien las estructuras microanatómicas involucradas) que es difícil excluir cualquiera entre centenares de conjeturas sobre la forma en que los dispositivos electrónicos deben conectarse para imitar el trabajo del cerebro. Dada una computadora enorme, casi totalmente desconocida, como es el cerebro, con trillones de elementos activos conectados en forma inescrutable, un científico computacional al que se le pida que adivine su modo de funcionamiento sin otros indicios que la información de que ella genera sus eductos utilizando computaciones masivamente paralelas que involucran sólo unos pocos centenares de etapas seriales sucesivas de procesamiento, sólo puede sentir una mínima confianza en cualquier especulación que él o ella aventure. El problema no es que no pueda imaginar la forma en que las propiedades conocidas de las neuronas puedan servir para sustentar las funciones inteligentes; el problema es que quedan abiertas demasiadas líneas de especulación para que sea viable una elección definitiva entre ellas sin evidencia adicional. El trabajo del teórico consiste, en consecuencia, en cultivar cierta sensibilidad frente a los indicios disponibles en la creciente y confusa masa de datos que el laboratorio de neurociencia nos ha proporcionado. En los párrafos que siguen reviso unos pocos entre los indicios más útiles.

El registro directo de la actividad de neuronas singulares ha sido posible durante varias décadas, y mediante la correlación

entre insumos cerebrales sensoriales con registros de neuronas singulares se ha podido obtener un cuadro bastante crudo del trabajo de los sistemas sensoriales del cerebro, por lo menos en las etapas iniciales del procesamiento neuronal. Estas etapas parecen preparar los datos de entrada para los primeros actos de reconocimiento, todavía enteramente misteriosos. Los estudios de este tipo sugieren que ciertas estructuras generales son comunes a diversas modalidades sensoriales. En muchos casos, las neuronas que manipulan información generada por los sistemas sensorios primarios, los cuales poseen una configuración natural uni o bidimensional, parecen estar dispuestas en páginas bidimensionales sucesivas (ya sea en el córtex cerebral o en diversas estructuras menores debajo de los lóbulos corticales). A menudo, la disposición de las células en estas páginas parece reflejar la geometría natural, o en cualquier caso alguna dimensión informacional significante, de los datos mismos. Por ejemplo, las células que cumplimentan las primeras etapas del procesamiento de la imagen en el córtex visual están dispuestas *retinotípicamente*, es decir que están en una correspondencia continua relativamente precisa con la retina del ojo (o, lo que es lo mismo, con la geometría de las imágenes que caen en la retina). Las células dedicadas al análisis de las sensaciones táctiles detectadas en la piel están dispuestas según una correspondencia *somatotópica* mucho más tosca con regiones de la piel, mientras que las células en las primeras etapas del sistema auditivo están dispuestas *tonotípicamente*, es decir, de acuerdo con las frecuencias auditivas frente a las que reaccionan. Las células que reaccionan a las propiedades más sutiles de los estímulos que ingresan también están dispuestas en geometrías regulares. Por ejemplo, el ángulo de respuesta máxima para las células sensibles a la orientación del córtex visual rota sistemáticamente a medida que uno se mueve a través de las pequeñas regiones del córtex; las células con campos retinales correspondientes en el ojo derecho e izquierdo residen en franjas verticales delgadas del tejido cortical (columnas de dominio ocular) adyacentes a las demás, pero nítidamente distinguidas entre sí. Presumiblemente, estas disposiciones celulares facilitan el intercambio de información que se necesita para detectar rasgos significantes en las corrientes sensoriales que ingresan, subrayando la alta intensidad y/o los cambios de colores, las orientaciones de los ejes o las esquinas agudas, por ejemplo. El

panorama sugerido por la evidencia disponible es un panorama de transformaciones sucesivas de estructuras de datos parecidas a imágenes (uni o bidimensionales) que producen estructuras secundarias parecidas a imágenes en las que los rasgos de estímulo que son potencialmente útiles para la formación de respuestas de más alto nivel se han hecho explícitos: una forma de procesamiento que no es nada sorprendente en términos de la computación científica. Aun en el ámbito sensorial, no conocemos más que unas pocas transformaciones específicas entre las que sufren los flujos de entrada, pero sabemos lo suficiente para pensar estos datos y su procesamiento en términos geométricamente extendidos, similares a imágenes. Más allá de esas etapas de procesamiento iniciales, relativamente bien conocidas, se entra en tierra incógnita, en una tierra en la que hasta hoy se ha probado imposible correlacionar la actividad neuronal observada con cualquier propiedad específica de las propiedades del estímulo externo.

De la neurobiología provienen otras ideas, consistentes con la evidencia ya revisada, a partir de la consideración de los patrones en los que se conglomeran las células del cerebro. Las neuronas, como las células que constituyen otros tejidos, son inicialmente movibles, es decir, capaces de migrar de sus posiciones originales, usualmente mediante una forma de «caminar» lento, guiada por la adhesividad selectiva de una célula en migración en relación con los tejidos sobre los que se apoya. Esta movilidad celular bioquímicamente regulada juega un papel fundamental en el modelado de los tejidos y órganos del cuerpo durante el desarrollo embrional: las páginas de células que llegan a constituir esos tejidos se erigen en muchos casos mediante la migración colectiva de sus células constituyentes, en forma muy parecida a la de una gran tienda de circo que se puede erigir mediante el movimiento colectivo de mucha gente que camina alrededor y tira de sus bordes.

En las neuronas, sin embargo, patrones de movilidad similares trabajan en formas significativamente distintas. Después de las fases iniciales del desarrollo embrionario, en vez de que sea el cuerpo mismo de la célula el que migra, una neurona arroja proyecciones (lo que serán sus axones y dendritas), cuyos extremos poseen pequeñas unidades móviles conocidas como conos de crecimiento. Cada uno de esos conos posee alrededor de veinte «pies» (seudópodos) de algo así como un milésimo de milímetro de

diámetro y treinta veces eso de largo, que permiten al cono moverse sobre la superficie de cualquier tejido con el que entre en contacto. Los seudópodos se extienden aparentemente al azar desde el cono de crecimiento en que se originan hasta que tocan la superficie de algún tejido cercano. Se adhieren entonces a él con una fuerza que está determinada por las cadenas de azúcar o almidón ligadas a moléculas de proteína modificadas (glucoproteínas) que están presentes en las membranas celulares de dos células en contacto. Una vez que se hace contacto, los seudópodos se contraen, empujando el cono de crecimiento y el axón que se desarrolla detrás de este cono aparentemente en dirección a la mayor adhesividad, lo mismo que una mosca se ve forzada a ir a las partes más pegajosas del papel cazamoscas en que está atrapada. Aunque sin duda están involucradas otras fuerzas determinantes de la dirección, estos efectos adhesivos, determinados en gran parte merced al brillante trabajo de Gerald Edelman y sus colaboradores en la Rockefeller University, parecen ahora estar en la base no sólo de la definición de los patrones de interconexión en el sistema nervioso, sino también del desarrollo embriológico en general.

Los conos de crecimiento de las neuronas continúan su camino, aparentemente hasta que cada una contacta una célula de destino marcada con alguna sustancia química a la que los seudópodos son sensibles, punto en el cual alguna reacción enzimática desconocida destruye la capacidad de los seudópodos para seguir moviéndose. El cono de crecimiento se transforma entonces en una estructura parecida a una sinapsis, que posteriormente se desarrolla en una sinapsis madura.

Este esquema del desarrollo del sistema nervioso no sugiere que el patrón de conexiones formado en los complejos cerebros de los mamíferos sea totalmente específico, en el sentido de que crea conexiones perfectamente determinadas entre neuronas específicamente identificadas, como si las neuronas fueran transistores en un chip de silicio artificial y como si las conexiones entre ellas se formaran depositando metales en una forma precisa. Más bien, este esquema sugiere un sistema que quizá contiene centenares, miles o decenas de miles de subespecies de neuronas, tal vez distinguibles bioquímicamente y quizá diferentes en forma significativa en sus reacciones particulares ante los estímulos externos, pero quizás interconectadas con una especificidad relativa-

mente global. Las reglas de crecimiento que se aplican pueden solamente especificar, por ejemplo, que una neurona de un tipo particular, originada en cierta parte de una capa específica del cerebro, se conectará por medio de un axón y una sinapsis a cualquier neurona de algún segundo tipo que esté cerca de alguna otra posición en otra región del cerebro. Los mecanismos de crecimiento conocidos son suficientes para dar cuenta de estructuras que poseen este grado de especificidad, pero parece dudoso que ellas puedan producir las estructuras mucho más específicas que caracterizan la circuitería de las computadoras y que a menudo entran en los modelos del sistema neuronal de los pensadores especulativos provenientes de la computación.

En relación con esto vale la pena señalar que en la actualidad no sólo carecemos de un conocimiento detallado del patrón de interconexión del cerebro, sino de una comprensión general de la cuestión más fundamental y rudimentaria de la forma en que especies bioquímicamente distintas de neuronas habitan el cerebro. En parte por esta razón, los teóricos que proponen modelos abstractos del cerebro a menudo comienzan suponiendo que todas las neuronas son funcionalmente idénticas y de acuerdo con eso modelizan a las neuronas como simples elementos de umbral que emiten señales cada vez que la suma de los estímulos que reciben, menos la suma de sus estímulos inhibitorios, excede algún valor de umbral fijo o ajustable. Es como si un investigador enfrentado con el problema de analizar un sistema computacional inmenso, con una arquitectura interna totalmente desconocida, comenzara suponiendo que todos sus circuitos de chips integrados son idénticos, simplemente porque a primera vista parecen más o menos iguales y porque cualquier hipótesis más cercana a la verdad fuera demasiado desalentadora. El examen sumario de las poblaciones de neuronas cerebrales, sin embargo, las muestra tan diferentes entre sí como los arbustos de jardín difieren de los sequoias gigantes. Además, incluso células de morfología externa aparentemente idéntica pueden diferir bioquímicamente en formas que hacen que sus reacciones a patrones parecidos de estímulos externos sean ampliamente diferentes. En las próximas una o dos décadas, el uso sistemático de baterías cada vez más poderosas de anticuerpos monoclonales disponibles como reagentes bioquímicos ultrasensitivos disipará gran parte de nuestra actual ignorancia sobre las variedades de neuronas.

Sin embargo, mientras nuestra ignorancia subsista, la utilidad de la información neuroanatómica, aun de la más estándar, se halla comprometida, dado que lo que se busca es conocer la forma en que se interconectan las poblaciones de células informacionalmente significativas (y, presumiblemente, bioquímicamente distinguibles). En contraste, la información disponible concierne sólo a la forma en que se conectan las regiones del cerebro.

La muerte masiva de células que se manifiesta inmediatamente después del nacimiento confirma la impresión de que el cerebro está diseñado para funcionar correctamente aun si las neuronas que lo constituyen se ligan de formas que sólo son aproximadamente correctas. Es bien sabido que en los mamíferos recién nacidos alrededor del 15 por ciento de las neuronas presentes en el neonato mueren durante la infancia temprana. La evidencia sugiere que muchas de esas neuronas representan ya sea un sobrecrecimiento o neuronas que por alguna razón han formado conexiones inapropiadas y por lo tanto no pueden recibir los estímulos eléctricos o químicos que se necesitan para mantenerlas viables. Evidencia adicional que sugiere que las interconexiones en el cerebro no son enteramente específicas proviene de experimentos en los cuales las conexiones sinápticas establecidas por una población de neuronas se destruyen cortando los axones que conectan estas sinapsis con los cuerpos celulares en los que se originan. Habitualmente, ese corte hace que se ramifiquen las neuronas vecinas y que se formen en consecuencia conexiones sinápticas anormales entre neuronas que por lo común no se conectarían en la región cerebral afectada. Esta evidencia sugiere que en el desarrollo de las interconexiones entre neuronas interviene un mecanismo de crecimiento competitivo, y que las neuronas invaden espacio sináptico desocupado en forma parecida a la de las hierbas que tienden a invadir un campo inicialmente vacío; no es una situación que favorezca la idea de un cableado preciso, como el de las computadoras.

Estas consideraciones sugieren que el cerebro puede ser incapaz de utilizar los patrones de procesamiento de información que son los más efectivos para las computadoras artificiales, incluso para las computadoras paralelas más grandes. A menudo, los sistemas de computación artificiales pueden generar más efectivamente los resultados deseados (a veces con notable eficiencia) utilizando secuencias de pasos de procesamiento elemen-

tales cuidadosamente diseñadas y cordinadas. En ese procesamiento, los conjuntos de datos que se procesan se mueven a través de una especie de danza de cuadrillas estrechamente coordinada y masivamente paralela, durante la cual cada ítem de dato interactúa con todos los ítem que encuentra, de forma de dejar el educto deseado en su lugar cuando el procesamiento termina. Cualquier falla en la sincronización o en una operación local que combine dos operandos, genera una onda de errores y deja un resultado ininteligible. Esos procesos paralelos delicadamente balanceados sólo generan los resultados que se esperan, o para el caso cualquier resultado útil, si cada movimiento de cada uno de los miles de ítem de datos que están siendo procesados tiene lugar precisamente en el momento especificado para él, y si cada uno de los millones de operaciones aritméticas o lógicas involucradas trabaja perfectamente. La evidencia que he citado sugiere que los sistemas biológicos no están cableados de una forma lo suficientemente precisa como para soportar un estilo de procesamiento de información tan extremadamente delicado. En particular, no tenemos evidencia de que el sistema nervioso opere de otra manera que no sea la de una perfecta asincronía, de modo que ninguna forma de procesamiento de información que requiera una sincronización precisa o que se vuelva sustancialmente más barata en su ausencia es un candidato atractivo para su uso en los sistemas neuronales.

Las consideraciones evolutivas también sugieren que el cerebro no hace uso de patrones de procesamiento de información delicadamente balanceados de la clase que es tan común y efectiva en la práctica computacional. Por cierto, la evolución procede acumulando pequeños cambios al azar, cada uno de los cuales afecta típicamente a un detalle de una de los miles o decenas de miles de moléculas de proteína cuya interacción determina la biología y la función celular. Para que las presiones evolutivas lleven el cambio más lejos, cada paso evolutivo debe proporcionar organismos que acarreen el cambio con suficiente ventaja como para favorecer su supervivencia, por lo menos marginalmente. Esta observación parece excluir los grandes saltos cualitativos que transforman un patrón de actividad establecido en otro patrón radicalmente distinto y delicadamente balanceado, como si se tratara de un programa compuesto de partes, ninguna de las cuales es útil hasta que la estructura total esté en su lugar. Los

algoritmos de procesamiento de información inteligentes requieren exactamente esas construcciones lógicas complejas e intervinculadas, otra razón para que su uso en un contexto biológico sea inapropiada.

Nuestro conocimiento de la neuroanatomía y la bioquímica de las neuronas, aún insuficientemente desarrollado, no nos proporciona información que nos permita modelizar específicamente la actividad de las neuronas en absoluto. En parte por esta razón, los teóricos han permanecido apegados a los modelos neuronales homogéneos y a teorías altamente conjeturales (aunque atractivas) respecto de que el cerebro o partes importantes de él progresan —desde una condición inicial informacionalmente blanca hacia un estado en el cual se ha codificado mucha información útil— mediante un proceso de aprendizaje que actúa a nivel sináptico. La más común entre las teorías de este tipo es una que Hebb propuso inicialmente en la década de 1940. De acuerdo con Hebb, la sinapsis que recibe estímulos excitatorios durante períodos en los que la neurona a la que se aferra está activada, se vuelve más sensitiva y por lo tanto actúa con más fuerza en ocasiones subsiguientes para estimular el disparo de la misma neurona. La eficacia de las sinapsis que no están involucradas en un patrón de estímulos sinápticos que causan repetidamente que se dispare una célula puede disminuir entonces en términos relativos, y quizá también en términos absolutos; con el tiempo, esas sinapsis se vuelven parcial o totalmente incapaces de estimular su célula.

Los mecanismos propuestos por Hebb permiten que células inicialmente indiferenciadas se vuelvan selectivamente condicionadas hacia una variedad de patrones que se pueden originar directamente en los sistemas sensoriales o indirectamente en las etapas iniciales del procesamiento neuronal. Su hipótesis ha ejercido influjo sobre los modelizadores neurales teoréticamente orientados, dado que no se halla en conflicto con ninguna evidencia actualmente disponible, aunque sugiere una forma en que el aprendizaje puede moldear las estructuras neurales que necesita muy pocos supuestos. Más aún, es fácil imaginar mecanismos bioquímicos, compatibles con la hipótesis de Hebb, que puedan permitir el despliegue de capacidades de procesamiento de información muy poderosas, tales como las formas generales de la memoria asociativa.

A despecho de este atractivo, sólo recientemente hemos comenzado a tener evidencia experimental tangible en sustento de la conjetura de Hebb, y esto sólo para unas pocas regiones del cerebro, y más notoriamente el cerebelo. La investigación reciente sobre esta importante estructura cerebral muestra que funciona, por lo menos en una pequeña parte, como un mecanismo para el almacenamiento de simples reflejos condicionados. Esta función se ha comprobado mostrando que una estimulación adecuadamente estructurada, simultánea e intensa de las neuronas apropiadas (específicamente, de las fibras cerebelares «trepadoras» y las células paralelas que se originan en las células granulares del cerebelo) causan cambios perdurables en la sensibilidad de las grandes células de Purkinje. Estas son las mismas células que presumiblemente participan en la formación de los reflejos condicionados pavlovianos simples. Es posible de este modo establecer un reflejo condicionado (por ejemplo, condicionando a un estímulo auditivo precedente el reflejo primitivo de parpadeo que se dispara cuando se sopla sobre la córnea), incluso si uno de los factores experimentales, o ambos (el soplido o el estímulo auditivo), que normalmente entra en su formación se reemplaza por la estimulación eléctrica directa o el correspondiente insumo cerebelar. Las modificaciones que se presentan durante estos condicionamientos artificiales eléctricamente inducidos se pueden localizar en una sola clase de sinapsis, a saber, las sinapsis entre las fibras paralelas y las capas múltiples de células de Purkinje que esas fibras atraviesan. Este trabajo experimental certifica brillantemente las conjeturas teoréticas concernientes al papel del cerebelo que David Marr y James Albus asentaron hace años. Estas conjeturas se inspiraron en el sorprendente parecido abstracto entre la microanatomía cerebelar y la disposición física de ciertos tipos de memoria de computadora.

Más allá de esta comprensión profundamente intrigante, pero todavía limitada, las teorías sobre el origen de las funciones neurales basadas en el aprendizaje permanecen sujetas a la objeción de que aún no sabemos nada sobre el lugar o el mecanismo de otros almacenamientos de memoria en el cerebro, y aún menos conocemos la forma en que se modifica la memoria para lograr el aprendizaje abstracto. Pese a que está muy difundida la creencia de que las sinapsis representan los lugares elementales del almacenamiento de recuerdos y que ese almacenamiento de alguna manera se realiza modificando la reactividad sináptica, no hemos sido aún capaces de

desarrollar suficiente evidencia bioquímica para sustentar esa creencia. Por ejemplo, ciertos estudios a menudo citados indican simplemente que ratas experimentales criadas en ambientes estimulantes desarrollan en apariencia números más grandes de sinapsis que ratas aprisionadas en ambientes sin estímulos. Además, los procesos de modificación sináptica revelados en estudios de sistemas nerviosos más simples (los famosos trabajos de Eric Kandel sobre la *Aplysia*, por ejemplo) no son específicamente hebbianos. Los cambios sinápticos determinantes que se ven en esas investigaciones parecen ocurrir en el lado transmisor (presináptico) de las sinapsis, más que en el lado receptor (post-sináptico) de las sinapsis, y de allí que no estén de acuerdo con los mecanismos supuestos por Hebb. De este modo, los teóricos que toman algunas hipótesis sobre el aprendizaje como su punto de partida escogen comenzar en un área de la neurociencia particularmente oscura.

Para funcionar efectivamente, los teóricos con experiencia de investigación en ciencia computacional e inteligencia artificial necesitan extraer una noción apropiada de neurocompatibilidad de la difusa masa de evidencia que viene del húmedo laboratorio de la neurociencia. Esta noción, simultáneamente, debe reflejar todo conocimiento detallado de la función del sistema nervioso central que arroje luz sobre las actividades de procesamiento de información del sistema nervioso y debe definir las restricciones sobre los modos de procesamiento neuronal que puedan orientar mejor los intentos del teórico para averiguar qué está pasando. Hasta ahora, éstos son los indicios que parecen más útiles:

1. El sistema nervioso debe hacer uso de algoritmos altamente paralelos que involucran sólo muy pocas etapas sucesivas de transformación de las corrientes de datos que ingresan.

2. Las etapas mejor conocidas del procesamiento sensorial inicial parecen involucrar transformaciones sucesivas de estructuras de datos parecidas a imágenes para subrayar rasgos que son probablemente los más importantes para la formación subsiguiente de respuestas de alto nivel. La configuración de estos datos en las páginas neuronales que los procesan está a menudo en correspondencia con parámetros continuos y variables, inherentes a los datos que se procesan (la posición retinal o la orientación del eje en el caso del ojo, por ejemplo, y la altura en el caso del sistema auditivo).

3. En el sistema sensorial se observan neuronas que difieren en la información que extraen de los datos que ingresan. Su existencia puede apuntar a la existencia de subespecies de neuronas morfológicamente parecidas, pero bioquímicamente distintas, dentro de áreas locales del tejido, lo que puede servir para transportar dimensiones separadas de un flujo de información que está ingresando. El número de esas subespecies informacionalmente significativas se desconoce, y posiblemente sea elevado.

4. La tosca pintura de las neuronas como dispositivos que suman señales excitatorias e inhibitorias y que pasan tanto de esa suma como lo que excede un umbral inherente, necesita refinarse para dar cuenta de complejas lagunas temporales y efectos no lineales, fácilmente resueltos por los complejos biociclos internos de todas las células, neuronas incluidas.

5. El nivel de adecuación de cableado en el sistema nervioso parece ser bajo, y parece no hacer ningún uso de pasos de procesamiento que involucren una sincronización fuerte de movimientos de datos o patrones de interconexión demasiado artificiales. Las formas de procesamiento que surgirían naturalmente en poblaciones de neuronas, posiblemente consistentes en múltiples subespecies interconectadas conforme a simples reglas de crecimiento, son más atractivas como conjeturas.

Máquinas de redes neurales

Aparte de reflejar el deseo de dar asistencia teórica a los neurocientíficos experimentales en su búsqueda de la forma en que trabaja el cerebro viviente, el creciente compromiso de los computadores científicos con la neurociencia tiene un segundo motivo. Este es el de usar el conocimiento del cerebro para que guíe el diseño de nuevas computadoras masivamente paralelas, las así llamadas máquinas de redes neurales. Aunque no es probable que produzcan resultados rápida y fácilmente, las contribuciones de la ciencia computacional a la neurociencia alcanzarán plena legitimidad científica. Que la neurociencia actual oriente el diseño de computadoras a corto plazo parece mucho más dudoso. Una lista sustancial de argumentaciones sustenta lo que afirmamos:

1. Aun con respecto a los sistemas sensoriales mejor conocidos, poco se sabe aún sobre el trabajo en detalle del cerebro. De las funciones cerebrales, aparte de los sistemas sensoriales, no sabemos esencialmente nada. De aquí que cualquier afirmación de que la arquitectura de una computadora específica imita al sistema neuronal sea pura conjetura.

2. Como tecnologías, el sistema nervioso viviente y las redes estructuradas de silicio que constituyen las computadoras son muy diferentes. El sistema nervioso es tridimensional y no sincronizado; probablemente debe tolerar altos grados de cableado defectuoso, pero puede formar decenas o cientos de miles de conexiones a cada uno de sus elementos computacionales, las neuronas. Al menos por el momento, los patrones de circuitos electrónicos están restringidos a las superficies bidimensionales de chips de silicio y también (excepto donde se utilizan patrones especiales, sumamente regulares) a unas pocas decenas de conexiones por elemento activo y unos pocos cientos por chip; estos circuitos pueden, sin embargo, cablearse con precisión casi absoluta de modo que puedan operar en estrecha sincronía.

3. Las computadoras pueden explotar cualquier patrón artificial de interconexión de hardware o de procesamiento de software que el trabajo intelectual de la máquina y los diseñadores de algoritmos traigan a la luz. Hemos afirmado que sólo una fracción minúscula de esos patrones de procesamiento está disponible para la actividad ligada a la evolución del cerebro viviente. Es revelador advertir que todos los proyectos importantes para diseñar y construir grandes máquinas paralelas hacen uso de estructuras altamente artificiales para la comunicación y el procesamiento. Esta afirmación se aplica también a la Máquina de Conexión de la Thinking Machines Corporation (comunicación en hipercubo y en matriz rectangular), al Procesador Masivamente Paralelo de la NASA, al Procesador Digital de Matrices de ICL (matriz rectangular), al Hipercubo de Intel Corporation (comunicación en hipercubo), al RP3 de IBM y a la Ultracomputadora de la New York University (comunicación en red omega). Más aún, en estas máquinas las computaciones utilizan algoritmos paralelos altamente artificiales y eficientes, y no procedimientos que sugieran las constricciones que afectan al procesamiento de la información en las estructuras de neuronas naturales.

De esta forma, las discusiones entusiastas que vislumbran un vasto potencial para algunas formas oscuramente caracterizadas de máquinas de redes neurales (y especialmente las propuestas para construir esas máquinas) parecen sospechosas. En cualquier instancia, todavía no ha aparecido ningún argumento serio que justifique esas afirmaciones. Las dificultades encontradas hasta ahora en la investigación son por cierto muy poco alentadoras.

Sin embargo, debe hacerse una excepción a esta reserva sobre la perspectiva de analogías neuronales en el diseño de dispositivos electrónicos, en favor del notable trabajo sobre sensores electrónicos integrados hecho por Carver Mead en el California Institute of Technology. La idea de Mead puede exponerse como sigue. Aunque, cuando se lo utiliza en la circuitería digital, un solo transistor sólo cumple las operaciones más elementales de la lógica booleana (de modo que, por ejemplo, se requieren varias docenas de transistores para implementar operaciones tan simples como la suma de dos dígitos decimales), mucho más sería posible si el mismo transistor se usara en forma analógica, y no digital. El uso analógico de la circuitería trata los voltajes del circuito como representaciones de valores numéricos, con una exactitud de uno o varios dígitos decimales; el uso digital asigna umbrales a esos voltajes clasificándolos como altos o bajos. La estrategia digital se ha vuelto ampliamente popular debido a que mejora decisivamente la estabilidad de las computaciones electrónicas y simplifica de algún modo la fabricación de circuitos, pero la pérdida de información y de velocidad potencial de procesamiento es sustancial. Si se los usa en forma analógica, unos pocos transistores pueden realizar operaciones matemáticas tan complejas como la multiplicación o la extracción de logaritmos, al menos aproximadamente; hecha digitalmente, la misma operación requeriría cientos de transistores.

Aunque esta ventaja potencial de la computación analógica ha sido bien comprendida por décadas, los sistemas analógicos han perdido terreno en forma constante frente a sus competidores digitales. En primer lugar, la precisión de los sistemas digitales se puede extender rápidamente a cualquier nivel que se desee simplemente agregando tantos dígitos como se quiera a la representación de una cantidad numérica. Sólo se requieren componentes estándar de costo fijo. En contraste, la precisión de los

sistemas analógicos se halla limitada inherentemente por la exactitud con que se pueden fabricar y aislar sus dispositivos componentes de perturbaciones físicas exteriores, tales como cambios de temperatura. Una consecuencia de ello es que los costos de los dispositivos analógicos escalan rápidamente con cada dígito de precisión adicional requerido y pronto alcanzan el límite de la inviabilidad absoluta.

Una segunda ventaja de los sistemas digitales es que pueden retener información con perfecta exactitud por períodos indefinidamente prolongados almacenándolos en dispositivos de estabilidad esencialmente perfecta: en las «memorias» de computadora que ahora son lugar común. Dado que la información analógica (valores de voltaje, por ejemplo) inevitablemente degrada y oscila con el tiempo, no se dispone de nada parecido en la esfera analógica. Las computadoras puramente analógicas no pueden, en consecuencia, almacenar sus programas en el mismo sentido en que pueden hacerlo las computadoras digitales. De allí que la compleja información de control para los sistemas analógicos, más las tablas extensivas de constantes o funciones auxiliares que sean necesarias, se deben almacenar digitalmente y convertir a forma analógica cuando se lo necesita para la computación analógica. Todavía peor, todos los datos intermedios se deben reconvertir a forma digital si se los ha de almacenar durante un tiempo. La desprolijidad y las limitaciones inherentes de esta situación han restringido la computación analógica a una esfera cada vez más angosta, hasta que en el presente se puede decir que la computación de este tipo en gran escala es casi inexistente.

La intuición de Mead expresa que hay un área importante en la que las desventajas de la computación analógica son irrelevantes, a saber: en el procesamiento de las corrientes de información sensorial, como la información de audio o las imágenes móviles. Aquí la precisión de los sistemas digitales es de escasa ventaja, dado que en cada caso es necesaria la conversión a partir de alguna forma sensorial analógica cruda, lo cual implica que los datos que alimentan un sistema de análisis, sea digital o analógico, necesariamente serán de precisión limitada. Más aún, muchos de los procedimientos comunes en el procesamiento inicial de estos datos, y en especial los que se encuentran en cualquier relación conjetural con el procesamiento sensorio inicial en el sistema nervioso, hacen poco o ningún uso de la información

previamente almacenada, de modo que la ausencia de memoria en los sistemas analógicos no es una objeción. En consecuencia, hay razón para esperar que las redes analógicas puedan procesar datos sensoriales de una manera que se beneficiará de su gran simplicidad y de su naturaleza compacta en relación con sistemas digitales comparables. Mead ha construido dos sistemas interesantes que hacen eso: un analizador de espectros de sonido modelado sobre la membrana coclear del oído interno y un detector óptico de movimiento cuya estructura es similar a la neuroanatomía retinal del ojo. Su trabajo puede sugerir muchas otras aplicaciones que permitan la combinación de la ventaja en performance de la computación analógica con el empaquetado extremadamente sofisticado y de ultra-alta densidad que soporta la actual tecnología de integración en gran escala (VLSI). Eso puede inspirar muchas imitaciones y abre nuevas direcciones en el diseño electrónico.

Sin embargo, en mi opinión, el trabajo de Mead es interesante como VLSI analógico antes que como neurociencia en silicio; en particular, su analizador de espectro sonoro modeliza la estructura mecánica del oído interno antes que las estructuras neuronales que reciben sus eductos.

Perspectiva

Pueden esperarse nuevos descubrimientos sorprendentes de la neurociencia experimental en la próxima década. Los éxitos extraordinarios de la biología molecular constituyen una base importante para esta afirmación optimista. Una vez que se sospecha la presencia de alguna proteína bioquímicamente importante (aunque al principio desconocida) en un tejido de interés, se pueden usar técnicas moleculares para producir cantidades sustanciales de anticuerpos para esa proteína. Una vez disponible («levantado», en la jerga del biólogo molecular) ese anticuerpo (que es simplemente una proteína que contiene una porción complementaria de algún detalle molecular de la proteína) detecta la presencia de su proteína con exquisita sensibilidad. Además, el anticuerpo se puede marcar radiactiva, magnética u ópticamente y se puede usar para hacer visible la proteína-blanco, o incluso rasgos microanatómicos particulares de esa célula bajo el micros-

copio electrónico. Por añadidura, se pueden cubrir las paredes de tubos de cristal con el anticuerpo, y esas columnas pueden usarse para concentrar la proteína en factores de un millón o más. Esa concentración abre un camino al análisis químico y estructural de la proteína, y de allí a la identificación de antagonistas bioquímicos de su actividad normal. Entonces, adosando tejido cerebral viviente a esos antagonistas, se puede paralizar la porción de funcionamiento normal que está mediada por la proteína y aislar su papel fisiológico específico y su relevancia para la actividad de procesamiento de información del cerebro.

A medida que se lleven adelante más y más comprensivamente investigaciones de esta clase para baterías enteras de proteínas significativas como receptoras de superficie en las neuronas, serán identificables poblaciones de células por las colecciones y concentraciones de moléculas receptoras en sus superficies. Además, llegaremos a saber la forma y velocidad con la que las neuronas responden a la activación de receptores de superficie particulares. Estas respuestas pueden ser rápidas y estar mediadas eléctricamente, o pueden ser lentas y activadas a través de largas cadenas de efectos bioquímicos intermedios, disparados por las moléculas receptoras iniciales.

Como se señaló antes, la identificación bioquímica de las subpoblaciones de neuronas residentes en el cerebro pondrán nuevamente en el candelero los trabajos de los neuroanatomistas, aumentando la relevancia que su laborioso seguimiento de las conexiones internas del cerebro tiene para nuestra comprensión del procesamiento de la información cerebral. Los estudios embriológicos de la especialización y la migración de poblaciones de células bioquímicamente identificadas dentro del cerebro en desarrollo revelarán los mecanismos que rigen la formación de esos patrones de conexión y mejorarán nuestra comprensión de los esquemas de procesamiento de información que usan las subestructuras del cerebro y de su complejidad. Nuevas técnicas mediadas química y ópticamente ya están comenzando a mejorar nuestra habilidad para observar en forma directa la actividad eléctrica del cerebro. Ahora, por ejemplo, podemos registrar simultáneamente la actividad eléctrica de varios cientos de células interconectadas. En algún punto podremos comprender los mecanismos bioquímicos específicos (o quizá los diversos mecanismos) que subyacen a la memoria, y eso nos permitirá formular

modelos mucho más específicos de los procesos de la memoria, sean del tipo de Hebb o no.

Aunque estos masivos esfuerzos experimentales involucrarán a miles de bioquímicos y neurocientíficos durante muchos años, podemos esperar que investigaciones como ésas, así como a su tiempo técnicas completamente nuevas descubrirán toda una masa de detallles concernientes a la función del cerebro. A medida que esta información salga a la superficie, la aspiración actual de los computadores científicos de integrar su conocimiento con el de los neurocientíficos aumentará en relevancia. Aquellos que en la comunidad de la computación científica han pagado lo suyo a la neurociencia experimental, digiriendo y siguiendo el rastro de una inmensa masa de información, podrán jugar entonces una parte importante en la extracción de amplios principios sistematizadores a partir de una selva inicial de detalles experimentales. Estas seguramente serán vislumbres que se hallarán en el mismo pináculo de la ciencia.

7
Cerebros reales e inteligencia artificial

George N. Reeke (h) y Gerald M. Edelman

La inteligencia artificial es una ciencia que se encuentra en una posición epistemológica similar a la de la medicina dental aristotélica. Aristóteles afirmaba que las mujeres tenían menos dientes que los hombres[1] y atribuía esta característica a que las necesidades de las mujeres eran supuestamente menores, por ser los hombres más fuertes y más coléricos;[2] pero nunca se atrevió a mirar en la boca de la señora Aristóteles para verificar su teoría. En forma parecida, la IA se ha desarrollado como una empresa casi enteramente *sintética*, bastante aislada del estudio *analítico* complementario de la biología de la inteligencia natural, representado por la psicología y las neurociencias. Para un biólogo, la estrategia de la IA frente al estudio de la inteligencia parece una extraña forma de comprender el cerebro, el cual se encuentra, después de todo, en la base de la inteligencia humana. Sin embargo, los biólogos y los computadores científicos comparten en general la perspectiva monística de que los sucesos mentales, incluyendo las manifestaciones de la inteligencia, reflejan necesariamente la actividad de las neuronas en el cerebro.

Es importante, entonces, preguntarse si los objetivos de la IA son realmente tan distintos de los de la neurobiología como para requerir métodos enteramente distintos de investigación y demostración. Porque si no lo son, hay mucho que ganar si se unen los experimentos con la teoría y el modelado sintético para construir una ciencia global del cerebro.

George N. Reeke (h.). Profesor asociado de biología molecular y evolutiva en la Universidad Rockefeller.

Gerald M. Edelman. Profesor Vincent Astor en la Universidad Rockefeller y director del Instituto de Neurociencias.

Comenzaremos argumentando que los objetivos últimos de la IA y la neurociencia son bastante similares, pero que se han oscurecido merced a supuestos epistemológicos erróneos, tomados por una parte de los argumentos de Alan Turing y Alonzo Church sobre las capacidades universales de resolución de problemas propias de las computadoras (sugiriendo que el cerebro debe comprenderse como si fuera una computadora) y por la otra del reduccionismo de la biología molecular (sugiriendo que se debe comprender el cerebro como una colección de unidades que intercambian señales químicas). Estos supuestos confinadores, junto a inmensas dificultades prácticas y experimentales, han mantenido muy ocupados a los partidarios de ambas estrategias, situando siempre los objetivos muy lejos de su alcance. De hecho, la consideración de la magnitud del problema con la debida modestia, sugiere que la percepción por sí sola es suficientemente difícil de comprender sin intentar saltar directamente de la percepción al aprendizaje, del aprendizaje a la transmisión social y al lenguaje, y de allí a las riquezas de la etología. En el presente, todavía está en pie el desafío de comprender cómo puede incluso moverse un animal, y será bueno para la IA tener en cuenta primero estas cuestiones fundamentales.

Como biólogos (es decir, como evolucionistas) que buscan comprender lo que percibimos casi como una negación dogmática de nuestra ciencia —al menos en los años recientes— por quienes intentan crear IA, comenzaremos preguntando qué es lo que intenta lograr la IA, y cuáles son los supuestos básicos sobre la naturaleza de la solución que engendran los paradigmas estándar de la IA. Discutiremos algunos problemas que han surgido en la aplicación de esos paradigmas y algunos intentos recientes por superarlos mediante lo que llamamos «miradas de costado hacia la biología», o «la aproximación de los físicos a las redes neuronales».

Estas nuevas estrategias, a pesar del engañoso rótulo de «computación con redes neuronales», toman su inspiración de la física estadística y la ingeniería, no de la biología. Resultan enormemente atractivas para la comunidad de la IA por diversas razones. Proporcionan una receta práctica para parcelar las computaciones de IA entre un número elevado de procesadores simples, con un crecimiento potencialmente enorme de la velocidad de computación. El paralelismo ha demostrado ser difícil de aplicar a la IA, aunque ya ha surgido en la computación numérica como la única forma indefinidamente extensible de superar el «cuello de botella de von

Neumann» (los límites fundamentales impuestos a la velocidad de los procesadores individuales por las leyes de la física, por ejemplo, dado que las señales no pueden viajar de una parte a otra de la computadora más rápido que la velocidad de la luz). En virtud de ser imitaciones de sistemas estadísticos, las nuevas estrategias también proporcionan aproximaciones estadísticas a problemas de optimización que se han mostrado recalcitrantes a todo intento de computación exacta.

No obstante, estas estrategias estadísticas comparten con la corriente principal de la IA la noción implícita de que los objetos y sucesos, las categorías y la lógica están dados y que la naturaleza del trabajo cerebral es procesar información sobre el mundo con algoritmos que conduzcan a conclusiones que lleven a la conducta. Este paradigma de procesamiento de la información no puede abordar ciertos problemas básicos concernientes a la naturaleza de la información y a la forma en que llegan a existir sistemas capaces de analizar señales portadoras de información. El supuesto de que las categorías y las señales que codifican información sobre ellas son la sustancia básica en torno de la cual se organizan las computaciones inclina constitucionalmente a la IA a identificar la performance perceptiva e intelectual con algoritmos. Este «problema categórico» conduce directamente a la inhabilidad de la IA para lidiar con la complejidad e impredecibilidad del mundo real.

Nuestro objetivo en este ensayo es señalar la naturaleza fundamental de este problema. Caracterizando las formas observables de inteligencia en el mundo biológico (entre otras, la inteligencia humana), presentaremos una estrategia basada en los principios biológicos más básicos, a saber, la teoría darviniana de la selección natural. Presentaremos una teoría que sugiere la forma en que la selección puede proporcionar la solución al problema categorial y la forma en que puede manifestarse en el sistema nervioso.[3] Describiremos diversos autómatas que desarrollan tareas que involucran categorización perceptual mediante mecanismos selectivos. Finalmente, discutiremos algunas formas en que este principio puramente biológico puede contribuir al progreso ulterior de la IA.

La naturaleza del proyecto de la IA

Inteligencia artificial es uno de esos términos que tienen un significado tan autoevidente que rara vez se lo define con cuidado.

Como resultado de esto, ha llegado a tener una interpretación tan amplia que agrega confusión al debate sobre sus méritos y perspectivas. Algunos han llegado tan lejos como para definir a la IA como «cualquier cosa en la que trabaje la gente de IA». Sin embargo, necesitamos una definición más significativa y no recursiva. Para ayudarnos a ver más claramente estas cuestiones fundamentales, excluiremos las empresas que poseen fines de ingeniería pura, es decir, los proyectos que intentan desarrollar soluciones computacionales efectivas a problemas que ya están bien comprendidos en principio. (Incluimos en esta categoría, esencialmente, todo el trabajo sobre los así llamados sistemas expertos y la programación lógica.)

En su influyente manual *Artificial Intelligence*, Patrick Winston define la IA como «el estudio de las ideas que permiten a las computadoras ser inteligentes». En lo que hace a los objetivos de este estudio, él afirma (en este orden) que «uno de los objetivos centrales de la IA es hacer más útiles a las computadoras. Otro objetivo central es comprender los principios que hacen a la inteligencia posible».[4] Los neurobiólogos considerarían la mención de estos dos objetivos en el mismo discurso un poco carente de equilibrio. Reconocerían la gran importancia práctica del primero, pero considerarían que tiene poco interés fundamental. Estarían de acuerdo, por otro lado, en que el segundo objetivo captura muy bien la naturaleza esencial de su propia empresa, aunque considerarían que esta formulación es más bien abstracta. En contraste, los neurobiólogos casi ciertamente se referirían a su propio interés en el descubrimiento y la validación de tales principios mediante la observación de los sistemas inteligentes. Por ahora, tales sistemas permanecen confinados al mundo de los organismos biológicos.

Así, como lo he sugerido, los objetivos principales de la IA y de la neurociencia son por cierto similares. La IA, sin embargo, arranca con una estrategia más formal hacia esos objetivos que la apartan del estudio de los sistemas biológicos «desaliñados». ¿Cuáles son los temas de investigación que la IA considera apropiados para la consecución de sus objetivos, y cuáles son los paradigmas de investigación que esos temas engendran? En 1961 Marvin Minsky presentó una lista que la mayoría de los actuales trabajadores en IA probablemente seguirían aceptando hoy: encontrar procedimientos efectivos para la búsqueda, el reconocimiento de patrones, el aprendizaje, el planeamiento y la induc-

ción.[5] Quizás ahora se agregaría como categoría separada la comprensión del lenguaje; de acuerdo con Minsky, la «búsqueda» incluiría una variedad de problemas de optimización y razonamiento así como de recuperación de información codificada en la memoria, y «reconocimiento de patrones» tendría que abarcar claramente la descomposición de datos sensoriales en objetos componentes, así como la categorización de estos objetos y sucesos después que (o mientras) han sido reconocidos.

Esta elección de temas refleja los supuestos epistemológicos a los que antes aludimos. En el paradigma de la IA estándar, tal como lo presenta Winston, la clave para encontrar procedimientos poderosos que puedan resolver esos problemas radica en descubrir representaciones apropiadas de la información relevante. Una vez que se da una representación que «hace explícitas las cosas adecuadas y expone las restricciones naturales»,[6] es mucho más simple desarrollar procedimientos puramente computacionales para manipular la información, aun en su representación codificada, de modo de obtener las soluciones deseadas. Winston está en lo cierto. Una vez que disponemos de una representación apropiada, muchos problemas devienen susceptibles de una solución automática. En nuestra perspectiva, sin embargo, el problema que requiere inteligencia es el problema original de encontrar una representación. Situar este problema en el dominio del diseñador del sistema más que en el del sistema diseñado es una petición de principios y una reducción de la inteligencia a la manipulación de símbolos.

La línea de razonamientos que conduce a esta visión supersimplificada y mecánica de la inteligencia tiene orígenes ilustres. Se la puede remontar por lo menos a Pascal y a Leibniz, y asume su forma presente en los días iniciales de la computación moderna. El sustento más importante de la teoría de las computadoras digitales, el principio de universalidad de Turing, se ha desarrollado antes que existiesen máquinas que funcionaran; quizás ésta es la razón de por qué se alentó a la gente a pensar en términos muy abstractos y generales sobre las computadoras y sus capacidades. Las limitaciones muy reales de las actuales computadoras y programas de computación sólo se hicieron claras después, cuando se disipó la excitación inicial y el paradigma básico ya estaba bien establecido. Entonces se vieron estas limitaciones como algo puramente práctico, que podría solucionarse con el tiempo. Después de todo, quienes creyeran otra cosa no permanecerían en el campo para invertir sus

energías creativas en construir lo que consideraban un castillo de naipes. De esta manera, sólo ahora se está sospechando más abiertamente, como antes lo sospecharon algunos,[7] que la IA podía experimentar dificultades fundamentales, diferentes de las dificultades meramente prácticas. Es muy improbable que éste sea el caso con invenciones tales como la máquina de vapor, en las que la práctica precedió a la teoría y la guió por los canales más fructíferos.

La justificación básica para la IA, que criticaré detalladamente luego, reza esencialmente como sigue:

1. Los objetos y los sucesos existen en el mundo. Se puede reunir información sobre ellos mediante los sensores apropiados. El objetivo de los sistemas inteligentes es procesar o transformar esta información de modo de proporcionar la base para el «planeamiento» o la «inducción» de que habla Minsky.

2. Dada una representación en forma de hilera de símbolos, su manipulación puede llevarse a cabo mediante reglas puramente formales que no necesitan hacer referencia al significado de los símbolos. Estas reglas pueden expresarse como algoritmos.

3. Un algoritmo puede ser ejecutado por cualquier máquina universal de Turing.[8] La misma existencia de estas máquinas implica que el mecanismo particular de cualquiera de ellas no es importante. Lo importante para comprender la inteligencia son los algoritmos, no la ferretería con que se ejecuta. En particular, lo que el cerebro hace se puede describir mediante algoritmos.

4. La tesis de Church sugiere además que si existe un método consistente y finito para resolver un problema dado, luego existe un método que puede correr en una máquina de Turing y dar exactamente los mismos resultados.[9] Por consiguiente, al menos para los problemas que pueden ser resueltos consistentemente en tiempos finitos y especificados, una máquina de Turing es tan poderosa como cualquier otra entidad que pueda resolver el problema, incluido el cerebro.

5. Dado que la inconsistencia hace la ciencia imposible y el tiempo infinito requiere inmortalidad, sólo tiene sentido discutir problemas de la clase que puede ser resuelta mediante máquinas de Turing.

6. Una vez que se han identificado los requerimientos informacionales de tal problema y se ha presentado un algoritmo, el problema está comprendido en algún sentido. Se pueden encontrar otros algoritmos que sean más elegantes o más eficientes en algunas implementaciones, pero todos los algoritmos, incluidos los que usa el cerebro, están sujetos a los mismos requerimientos informacionales y se pueden comprender de la misma forma.

7. Por lo tanto, el cerebro es equivalente a una computadora, o al menos la computadora es un modelo adecuado de las cosas interesantes que hace el cerebro.

El paradigma estándar para llevar adelante la investigación en IA, completamente sintético, se sigue inmediatamente de esta línea de argumentos: elegir un problema significativo en que todo el mundo esté de acuerdo en que requiere para su solución el uso de inteligencia; identificar los elementos de información que se necesitan para lograr una solución del problema; determinar cómo podría representarse esta información en una computadora; encontrar un algoritmo que pueda manipular esta información para solucionar el problema; escribir un código computacional que implemente ese algoritmo y ponerlo a prueba contra instancias muestreadas (y usualmente también simples) del problema.

Esta estrategia ha conducido a un cierto número de impresionantes demostraciones. Por razones que no son accidentales, las más exitosas de ellas han sobrevenido en áreas que satisfacen más obviamente las condiciones para aplicar la tesis de Church: la resolución lógica de problemas y la demostración de teoremas (GPS, MACSYMA y el lenguaje Prolog, por ejemplo), identificación de objetos en imágenes (ACRONYM, CONSIGHT), jugar al ajedrez y otros juegos (CHESS, KAISSA, BELLE, PARADISE), comprensión del lenguaje humano en dominios limitados (SHRDLU, BORIS) y sistemas expertos, que combinan inferencia basada en reglas y técnicas de interface de lenguaje natural con bases de datos específicas de ciertos dominios (MYCIN, DENDRAL, PROSPECTOR, XCON). Sin embargo, todos estos programas comparten una cualidad común que John McCarthy[10] y otros han señalado repetidas veces: son «frágiles» en el sentido de que, presionados contra los bordes, tienden a «quebrarse». En otras palabras, los programas carecen de conocimiento y razonamiento de sentido

común, no «conocen» sus propias limitaciones. Son insensibles al contexto y tienden a dar respuestas sumamente incorrectas a preguntas que están ligeramente fuera de los dominios para los que fueron programados. Sus respuestas son consecuencias perfectamente lógicas de las reglas construidas dentro del sistema, pero al observador equipado con facultades humanas normales de razonamiento le pueden parecer arbitrarias e incluso misteriosas.

No es fácil encontrar una solución general a este problema de la fragilidad. Pese a que puede parecer sólo otra instancia de la dificultad general de establecer y mantener grandes sistemas de programas, parece que existe una diferencia cualitativa, que se origina en la apertura del lenguaje natural y en la necesidad de experiencia en el mundo real para adquirir un repertorio competente de sentido común. En la próxima sección veremos algunas soluciones a este problema que se han intentado en el paradigma de la IA estándar. Luego explicaremos por qué creemos que el problema está en el paradigma mismo y cómo se puede encontrar la solución mirando al sistema nervioso y a las formas en que éste trata con la complejidad y la novedad del mundo.

Miradas laterales a la biología

Cuando se experimentan dificultades en una empresa importante, el primer impulso es hacer más de lo mismo o hacer lo mismo mejor. En el caso de la IA, este impulso ha tomado dos direcciones: mejorar el hardware y mejorar el software. Cada una de estas direcciones ha dado ímpetu al desarrollo de nuevos métodos que involucran la representación de información mediante los estados de procesadores simples y las conexiones entre ellos (los así llamados modelos conexionistas o de las redes neuronales).

Primero, el hardware mejorado ha hecho posible evaluar más condicionales (pruebas lógicas programadas) por segundo, almacenar bases de datos más grandes y examinar imágenes con mayor detalle que antes. Esta computación más rápida ciertamente condujo a una performance mejorada de los sistemas de IA tradicionales. Los autómatas de ajedrez, por ejemplo, pueden ahora «mirar más profundamente» (más movidas hacia adelante) en las posiciones en

el tablero.[11] En última instancia, sin embargo, la velocidad del hardware sólo puede evolucionar hasta cierto punto. A su momento, se deberá introducir procesamiento en paralelo para seguir progresando, pero el uso efectivo del procesamiento en paralelo requiere nuevas técnicas de software. Entre las nuevas técnicas más intrigantes desarrolladas para reemplazar el software tradicional de IA están los modelos reticulares, que son ideales para la implementación en sistemas de multiprocesadores cuyas estructuras físicas son comparables o se han hecho para imitar la estructura de red de los modelos.

Segundo, se ha proclamado ampliamente que las nuevas técnicas de software sirven para abordar este problema de la fragilidad. Los programas de computadora no procedimentales prometen relevar al programador de la responsabilidad de especificar la secuencia de pasos que se necesitan para resolver un problema en particular (en realidad, esos lenguajes sustituyen los métodos de secuenciación estándar que están preconstruidos en el compilador y eso, debido a su generalidad, es dudosamente óptimo ante problemas particulares). Estos lenguajes permiten que los programadores introduzcan grandes números de reglas específicas de un dominio en sistemas expertos antes que la complejidad devenga totalmente imposible de manejar. Se deben introducir algunas formas de razonamiento «no-monotónico»[12] para mitigar los problemas que surgen cuando se aplican esas reglas sin verificaciones de consistencia precisa antes de usar el programa. Sin embargo, aun esas técnicas requieren un vasto esfuerzo de ingeniería de software para cada problema atacado. La posibilidad de resolver algunos de estos problemas con métodos orientados a redes que reducen aún más la cantidad de programación explícita ha proporcionado un segundo ímpetu importante al desarrollo de los métodos de redes simuladas dentro de la IA. Curiosamente, sin embargo, estos desarrollos tienen lugar en un relativo aislamiento del trabajo anterior y contemporáneo de los biólogos que intenta hacer modelos que incorporen algo de la rica diversidad funcional del sistema nervioso real. (Algunos ejemplos de nuestro propio trabajo aparecerán más tarde en este artículo.)

Sólo seguiremos el rastro de las principales corrientes en el desarrollo de esas redes neurobiológicas; se pueden encontrar más detalles en las obras enumeradas en las notas. Aunque gran parte de esta actividad ha sido reciente, la noción de que se pueden emplear redes para ejecutar computaciones en realidad precede

a la introducción de las computadoras digitales programables. La más influyente entre las discusiones tempranas ha sido probablemente el artículo de Warren McCulloch y Walter Pitts,[13] en el que la actividad de las redes neuronales se identificó con las operaciones del cálculo proposicional. (Utilizamos el término *neuronal* cuando está implicada una identificación estrecha con las propiedades de las neuronas reales. Usamos el término *neural* cuando sólo está involucrada una similitud general con las neuronas reales.) Las simulaciones reales de autómatas de reconocimiento basados en redes fueron realizadas por Frank Rosenblatt antes de 1958,[14] pero las limitaciones teoréticas de sus «perceptrones» pronto fueron señaladas por Marvin Minsky y Seymour Papert,[15] y el interés en los modelos de redes se desvaneció hasta la reciente introducción de modelos no lineales más complicados que no comparten esas limitaciones. Otras líneas de investigación que han ejercido influencia sobre diversos modelos actuales incluyen redes que están especialmente dotadas para propósitos especiales, como el sistema de David Marr y Tomaso Poggio para computar la disparidad estereoscópica,[16] y las redes en las que los nodos se identifican con «unidades» e interconexiones cognitivas con relaciones entre conceptos, como en el modelo ACT de John R. Anderson para recuperar información de la memoria.[17]

De esta forma, el escenario estaba listo para la «estrategia de los físicos frente a las redes neuronales», a medida que muchos científicos percibían la necesidad de diseñar redes para las que se pudiera prever la convergencia y las propiedades del aprendizaje mediante teoremas matemáticos, más que por medio de largas ejecuciones en computadora. Los sistemas lineales proporcionaban un punto de partida relativamente tratable,[18] pero es esencial la no linealidad si han de distinguirse las categorías sin solapamiento. No importaba si la analogía tenía que ser bastante forzada. Para esos trabajadores, la posibilidad del análisis era más importante que si esos sistemas se relacionaban o no con el problema básico de la inteligencia. Todo lo que uno tenía que hacer era identificar nodos de la red con spins y las conexiones con las interacciones espinales, y uno podía hablar entonces de una «energía» que se incrementaba con el número de pares de espines incompatiblemente conectados. A partir de cualquier estado inicial, el sistema se «relajaría» hacia un estado de energía mínima, y esos estados podían identificarse con «memorias» o «recuerdos»

codificados en los coeficientes de conectividad de la red. Cada memoria tendría una «zona de atracción» en el «espacio de estados» circundante. Esta descripción ligaba claramente las redes neurales con los recientes avances en la dinámica de los sistemas no lineales y con la teoría del caos. Los modelos mejor conocidos de esta clase son los de John Hopfield y sus colaboradores, quienes se han ocupado primariamente de la construcción de redes listas para usar para varias tareas más que con algoritmos de aprendizaje para esas redes.[19] Este último elemento ha sido agregado en los modelos de «máquinas de Boltzmann», que toman su nombre de su uso del concepto de energía de la mecánica estadística asociada con el físico Ludwig Boltzmann, de fines del siglo diecinueve.[20] Se ha explorado celosamente la representación de conceptos en esas redes, y se han presentado aplicaciones ilustrativas.[21]

Aun con la computación paralela y con nuevos algoritmos de aprendizaje, sin embargo, la posibilidad de entrenar una red en un número suficiente de circunstancias para conferirle *sentido común* parece alejarse para siempre a la distancia a medida que se contemplan las excepciones a las excepciones que están presentes en las situaciones de la vida real. Esta composición inacabable de excepciones está cerca de revelar la verdadera naturaleza del problema de la fragilidad, que es que ningún monto de anticipación, planeamiento o programación puede siquiera enumerar, a priori, todas las variantes de una situación rutinaria que puede ocurrir en la vida cotidiana. Parecería, en lo que es quizás una analogía al clásico «problema del detenimiento» en el análisis de los algoritmos (el problema de determinar si un algoritmo determinado correrá para siempre o se detendrá alguna vez), que la única forma de determinar todas las respuestas que necesita un sistema para tratar con las vaguedades del mundo real es exponerlo al mundo y dejarlo «correr». De esta forma, cada sistema será diferente y fundamentalmente improgramable.

Decimos que los modelos conexionistas «miran de costado hacia la biología», porque ellos toman su inspiración y gran parte de su terminología de las redes neurales en los organismos vivientes; pero éstas no son modelos de redes neurales (ni intentan serlo). Los físicos, en su búsqueda de simplicidad, no están preparados para tratar con un sistema cuyo aspecto principal reside en la variabilidad más que en la regularidad. En el intento de encontrar regularidad en sistemas biológicos, se han introdu-

cido en la simulación de los sistemas conexionistas muchos rasgos que son muy poco biológicos. Estos incluyen la noción de la memoria como una réplica o transformación de «información» dada en el mundo (las memorias humanas son en gran medida sensitivas al contexto y al afecto, y en cierta extensión son no verídicas); la concepción de la recuperación de información en la memoria como el relajamiento de la red a un estado estable (un cerebro está expuesto continuamente a patrones de insumo cambiantes y no tiene oportunidad de congelarlos mientras espera acercarse al equilibrio); la idea de la minimización de la energía mediante la solidificación simulada [*simulated annealing*] (un cerebro decide las acciones más rápidamente de lo que pueden hacerlo los procedimientos de solidificación conocidos en los modelos de redes que operan a la velocidad de las neuronas reales); la noción de las conexiones singulares bidireccionales o simétricas (las conexiones sinápticas en el cerebro son monodireccionales); y la idea de que el aprendizaje puede proceder atornillando el producto de salida de un sistema en un valor deseado, mientras los pesos sinápticos se ajustan de acuerdo con alguna regla (el educto motor del cerebro, en general, no se puede imponer externamente). Alguno de esos elementos está presente en todos los modelos conexionistas.

Estos rasgos irreales deberían ser una señal de advertencia de que algo está severamente equivocado en los supuestos básicos detrás del paradigma de la IA, incluso cuando se los modifica mediante la introducción de procesamiento en paralelo en los sistemas inspirados por las redes neurales. La cualidad *ad hoc* de estos supuestos sugiere que el verdadero problema radica más profundamente que en los detalles de las simulaciones de redes; debe tener que ver con el concepto de información y la forma en que llega a representarse y transformarse en los sistemas inteligentes que han evolucionado en la naturaleza.

Asentada en el comienzo de la cadena de deducciones enumeradas en las páginas 172 y 173, que para la IA justifica la noción del cerebro como computadora, está el supuesto de que la información existe en el mundo; está justamente ahí, esperando ser manipulada. También está la idea de que el organismo es un *receptor* más que un *creador* de criterios que llevan a la información. Una vez que se concede la existencia previa de esa información externa, es enteramente natural proceder sin más rodeos al trabajo de programar las reglas para tratar con ella. En este punto,

el daño ya está hecho. Todos los esfuerzos por programar esas reglas caen rápidamente en cinco círculos viciosos que tienen en común la transferencia al programador de funciones que pertenecen propiamente al sistema putativamente inteligente:

1. *El problema de la codificación.* El programador debe encontrar una representación adecuada de la información para ponerla en forma adecuada a la manipulación simbólica. Habitualmente no está claro por anticipado qué manipulaciones simbólicas serán requeridas y cuáles han de ser los requerimientos antecedentes de la representación.

2. *El problema categorial.* El programador debe especificar un conjunto suficiente de reglas para definir todas las categorías con las que el programa debe tratar. Es difícil ver por anticipado cuáles de estas categorías deben existir en el mundo real, y mucho más definirlas.

3. *El problema del procedimiento.* El programador debe especificar por anticipado las acciones que debe tomar el sistema para todas las combinaciones de insumos que puedan ocurrir. El número de esas combinaciones es enorme y crece aún más cuando se tienen en cuenta los aspectos relevantes del contexto. La conducta de los organismos biológicos con sistemas nerviosos reales deviene casi por completo impredecible en tales circunstancias.

4. *El problema del homúnculo.* Se requieren mecanismos separados para interpretar las hileras de símbolos que produce cualquier sistema formal de procesamiento de información. Las hileras pueden no tener sentido dentro del sistema formal mismo. Pero entonces las propiedades necesarias de la inteligencia están corporizadas en el observador, no en el sistema. Para evitar una regresión infinita, el programador está obligado a especificar todos los procedimientos que debe seguir el observador.

5. *El problema del desarrollo.* ¿Puede llegar a existir un sistema programado sin un programador? Los sistemas biológicos inteligentes existen, aunque ellos han evolucionado y no han sido programados, ya sea como especies o como individuos. En consecuencia, el argumento de la IA de que el cerebro lleva a cabo computaciones como lo hacen las computadoras conduce a una contradicción: los cerebros deben

tener programas, pero al mismo tiempo no deben ser programados.

Sostenemos que la solución a estos laberintos puede encontrarse examinando las inteligencias naturales existentes y otros sistemas biológicos en un intento por comprender cómo es que llegan a existir y cómo es que operan sin programas previos. En el resto de este ensayo, afirmaremos que una forma de selección, afín a la selección darviniana, pero que opera sobre colecciones de neuronas en el cerebro de un organismo durante toda su vida, proporciona la única base firme para una teoría de la categorización y la inteligencia.[22] La evolución, extendiendo de esta forma la selección al cerebro individual, ha sido capaz de eliminar la necesidad del programador.

Mirando de frente a la biología

Del análisis precedente surge claramente que debe rechazarse la noción de que hay una información preexistente en el mundo. El requerimiento esencial para el aprendizaje, la lógica y las otras funciones mentales que son los temas habituales de la IA es la habilidad para categorizar objetos y sucesos sobre la base de las señales sensoriales que alcanzan el cerebro. La variedad de experiencias sensoriales es al mismo tiempo vasta y única para cada individuo. Las categorías mismas no están presentes en el ambiente, sino que deben ser preconstruidas por cada individuo de acuerdo con lo que resulte adaptativo para su especie y para su propia circunstancia particular. La especificación a priori de reglas para la categorización, aplicable a todos los individuos y a todos los contextos, queda anulada por la complejidad, variabilidad e impredecibilidad del mundo macroscópico. Para hacer las cosas peores, las categorías construidas por un organismo no se pueden fijar, sino que tienen que cambiar constantemente en respuesta a las nuevas experiencias y a las nuevas realidades en su parte del ambiente. La única forma en que pueden validarse las categorías construidas de modo individualista es volviéndose a remitir constantemente al mundo a través de la conducta. Sin embargo, una vez que se ha establecido este proceso de categorización adaptativo, el resto de la tarea —la construcción de categorías de orden más elevado, los recuerdos y las asociaciones— se simplifica enormemente.

Nuestra primera tarea, entonces, es la de construir una teoría satisfactoria que vaya más allá del procesamiento formal de información, llegando a considerar la forma en que llega a existir la información en un mundo sin etiquetas, qué relaciones existen entre las señales en el cerebro y las categorías que ha construido y cómo las interacciones de estas señales mantienen la conducta sin el beneficio de códigos preordenados para darles sentido o de algoritmos preordenados para procesarlas. El problema abarca tanto la memoria como las señales transitorias: ¿Cómo puede funcionar la memoria en ausencia de un almacenamiento replicativo como el que se encuentra en una computadora, y cómo se puede combinar la memoria con las señales sensoriales concretas para producir conductas que mejoren la supervivencia del organismo? Para construir tal teoría, debemos considerar todo el sistema biológico, sus orígenes evolutivos y su desarrollo como individuo, desde el embrión hasta el adulto en plenitud.

Vale la pena comenzar considerando algunos hechos básicos de la neurobiología. En primer lugar, los sistemas nerviosos están organizados como redes con distintas áreas que poseen diferentes patrones de conectividad, aparentemente especializados para diferentes funciones. A menudo estas redes están conectadas entre sí en una secuencia de mapas. Caracterizamos esta anatomía como una «heterarquía» (una disposición en la que las subredes están interconectadas en formas que no siguen una jerarquía estricta). En esta disposición, las subredes con diferentes funciones interactúan para sustentar funciones más complejas que las subredes no poseen por separado. Las sucesivas regiones parecen haberse agregado gradualmente en el curso de la evolución, cada una contribuyendo a una nueva función pero trabajando en conjunto con las regiones más viejas para conferir al cerebro un alto grado de redundancia funcional. No hay ningún análogo aparente a las funciones del reloj y del decodificador de instrucciones en una computadora. En lugar de eso, hay un alto grado de paralelismo en las operaciones de las regiones funcionales, así como en las respuestas de las neuronas individuales. Las superposiciones extensivas de arborizaciones dendríticas y axonales sugieren una degeneración funcional en la cual hay muchos caminos alternativos entre dos puntos cualesquiera en la red, incluso dentro de un solo mapa. Que ninguna neurona parezca ser indispensable para ninguna función sugiere que sólo los

patrones de respuesta sobre muchas neuronas pueden tener significación funcional. En la medida en que las neuronas poseen velocidad y rango dinámico limitados, probablemente no desarrollan algoritmos computacionales en nada que se parezca a la forma en que lo hace una computadora.

La última observación, la más elocuente, es la enorme diversidad de las poblaciones neuronales. Esta diversidad se percibe a todos los niveles filogenéticos pero es mayor en las formas más elevadas. La diversidad se extiende al número y disposición de las neuronas en animales genéticamente idénticos e incluso a la estructura puntual de neuronas individuales de función conocida en aquellos animales en los que esas estructuras pueden identificarse. Se puede calcular fácilmente que no hay suficiente información en el DNA para especificar solamente las ubicaciones de todas esas neuronas y sus conexiones respectivas. De esta forma, durante el desarrollo deben operar mecanismos indeterminados, dinámicos, epigenéticos (mecanismos que reflejan la influencia del entorno local sobre los programas genéticos que se desenvuelven en las células individuales) para determinar la estructura fina del sistema nervioso. Esto no es lo que uno esperaría si los sistemas nerviosos estuvieran optimizados por diseño para desarrollar funciones cognitivas específicas. Más aún, la variabilidad parece no ser sólo una consecuencia inevitable de procesos de desarrollo esenciales, sino un rasgo seleccionado evolutivamente. Mientras que esa variación en el cableado llevaría a la falla total en una computadora, parece cumplir un papel funcional en el cerebro. La naturaleza de este papel quedará clara cuando consideremos la forma en que actuarían los procesos selectivos para dar la categorización flexible y adaptativa que hemos dicho que es necesaria para que un animal pueda tratar con el mundo sin el beneficio de un programa preordenado.

Los hechos biológicos que hemos señalado, particularmente la variación estructural en todos los niveles del sistema nervioso, sugieren que la selección posee un papel no sólo en el desarrollo de los sistemas nerviosos sino también en su funcionamiento en la madurez. En los sistemas selectivos las unidades funcionales no están específicamente construidas para desarrollar sus funciones en forma óptima, sino que en lugar de eso se seleccionan a partir de conjuntos mucho más grandes que se llaman «repertorios». Las unidades componentes de un repertorio se constru-

yen con una amplia variedad estructural que es suficiente para cubrir, con superposiciones, el rango de funciones posibles necesarias en cualquier instanciación particular del sistema. La selección ocurre durante la experiencia sin ninguna alteración ulterior de las propiedades funcionales de las unidades ya construidas. Más bien, las unidades seleccionadas se multiplican o amplifican de tal manera como para contribuir más a las futuras respuestas del sistema que las unidades no seleccionadas. Los procesos selectivos de esta clase, con diferentes modos de amplificación apropiados a cada sistema, proporcionan los mecanismos básicos del sistema inmune y, por supuesto, de la evolución misma.

La consideración de la necesidad de que el sistema nervioso proporcione a los organismos adaptabilidad conductual para sobrevivir en un ambiente hostil sin conocimiento o programación anterior, la consideración de la variabilidad estructural encontrada en todos los sistemas nerviosos y otras consideraciones han sugerido la idea encarnada en la teoría de la selección del grupo neuronal (SGN)[23] de que el cerebro es de hecho un sistema selectivo que opera en tiempo somático (es decir, durante la vida de un organismo individual). El pensamiento de la población, el modo teorético fundamental de la biología en el cual se tienen en cuenta las propiedades de las poblaciones, así como las de los individuos, se introduce por consiguiente en la consideración de la forma en que trabajan los cerebros individuales.

La teoría de la selección del grupo neuronal

De acuerdo con la teoría de la SGN, hay dos clases de sucesos selectivos que juegan papeles críticos en la conformación del desarrollo del sistema nervioso. Durante la formación del cerebro del embrión, la selección entre células neuronales en competencia y sus procesos determinan la forma anatómica y los patrones de conectividad sináptica del sistema nervioso. Esta selección para la conectividad se elabora mediante mecanismos evolutivos de adhesión y movimiento de células, crecimiento diferencial, división celular y muerte de células. Dadas sus propiedades dinámicas, estos mecanismos selectivos introducen variación individual en las redes neuronales. Más tarde, durante la experiencia

posnatal, la selección entre diversos grupos de células preexistentes, cumplimentada por la modificación diferencial de fuerzas o eficacias sinápticas *sin cambios en el patrón de conectividad*, da forma al repertorio conductual del organismo de acuerdo con lo que posee para él valor adaptativo en su econicho.

Un sistema ha de poseer tres rasgos para ser selectivo: 1) debe tener un repertorio a priori de entidades variantes capaces de responder a estados relevantes del entorno; 2) los miembros individuales de este repertorio deben tener oportunidades extensivas de encontrar la rica diversidad del ambiente, proporcionando oportunidades para la selección y 3) el sistema debe poseer un mecanismo para amplificar diferencialmente las contribuciones relativas de aquellos miembros del repertorio que en algún sentido han sido favorecidos o seleccionados en sus interacciones con el entorno. De acuerdo con la teoría de la SGN, los repertorios en el sistema nervioso comprenden grupos de 50 a 10.000 neuronas, capaces, como resultado de sus interconexiones, de responder a patrones de actividad particulares que llegan a sus sinapsis. Estas interconexiones se forman durante el desarrollo, previo a toda experiencia. Los insumos a los que los grupos responden se originan en última instancia en los órganos sensoriales (el encuentro con el ambiente), pero con frecuencia son transmitidos antes por otros grupos neuronales. La modificación selectiva en las fuerzas de las conexiones sinápticas (amplificación diferencial) lleva a la compartición y estabilización de los circuitos en funcionamiento fuera de sus redes fijas. En las formas más elevadas, las respuestas de las poblaciones neuronales están significativamente influenciadas por similitudes entre constelaciones presentes y pasadas de señales sensoriales. Esta reevocación de respuestas previas constituye la base de lo que llamamos memoria. La memoria es una consecuencia de la amplificación selectiva, la cual conduce a una velocidad o fuerza creciente de las respuestas seleccionadas cuando se repiten patrones similares de estímulo.

Para ser capaz de responder adecuadamente a un amplio rango de insumos novedosos, un sistema selectivo debe tener un número suficiente de unidades en sus repertorios. Se puede derivar una relación entre los tamaños de los repertorios y las especificidades de los grupos individuales.[24] Si el reconocimiento es demasiado específico, el sistema fallará porque no habrá forma de situar grupos suficientes en un repertorio finito para reconocer todos los

estímulos posibles; de igual modo, si la especificidad es demasiado amplia, el sistema fallará porque pueden confundirse estímulos con diferencias significativas. Las especificidades han de ser entonces intermedias, permitiendo que diversos grupos respondan más o menos bien a un estímulo dado. Este fenómeno, que llamaremos «degeneración», es crítico para la comprensión de los sistemas de reconocimiento selectivo. La degeneración asegura que cualquier problema perceptual tenga suficientes soluciones potenciales. El contexto determina exactamente qué combinación de grupos responde a una situación dada, y en consecuencia qué solución se selecciona. La degeneración también asegura que se cubra el «espacio» entero de posibles estímulos y que el sistema tenga la redundancia funcional necesaria para hacerlo a prueba de fallas en caso de pérdida de grupos individuales.

Un concepto adicional que es crítico para la teoría de la SGN es el de «reentrada» o intercambio de grupos de salida, usualmente en forma de una distribución mapeada, que tiene lugar entre un repertorio y otro en una misma etapa o en una etapa anterior de procesamiento neuronal. La reentrada proporciona un mecanismo para correlacionar respuestas en posiciones correspondientes de mapas relacionados, de modo de asegurar la consistencia a lo largo de todo el sistema con respecto al estado concreto y a la continuidad espaciotemporal del entorno. La reentrada asegura que los subrepertorios a todos los niveles del sistema nervioso se mapeen entre sí y sobre el mundo exterior, obviando cualquier necesidad de conmutar de contexto, definiciones de tiempo o cualquier otro aparato de librería utilizado en las computadoras. La reentrada comprende a la retroalimentación pero es más general. Por ejemplo, dos repertorios diferentes en caminos paralelos independientes, cada uno de los cuales muestrea disyuntivamente diferentes aspectos de las señales, pueden clasificar los estímulos de acuerdo con diferentes criterios y se pueden interconectar a niveles más altos. Esos repertorios interactuantes forman «parejas de clasificación» que, por su interacción mutua, pueden ejecutar clasificaciones más complejas que las que cualquier repertorio puede ejecutar por separado. Otra forma de reentrada involucra al sistema total de la criatura más el entorno. En esta forma «global» de reentrada, el educto motor del organismo influye sobre los sistemas sensoriales cambiando la disposición relativa de los objetos o la posición del

organismo en el espacio. Los mapeos a todos los niveles del sistema nervioso se unen en un bucle global. La iteración de la actividad en este bucle puede alterar la reentrada en los mapas locales y conducir a conductas modificadas que son más adaptativas para el organismo.

La teoría de la SGN es consistente con los hechos biológicos que hemos resumido. Toma ventaja de la inevitable varianza en la conectividad que introducen los sucesos epigenéticos durante la construcción de redes neuronales para proporcionar un mecanismo plausible para la categorización sin descripciones preprogramadas, homúnculos o aprendizaje reforzado. La teoría se basa en principios de selección similares a los que gobiernan la evolución de las especies; los mecanismos, por supuesto, son distintos. Esta estrategia no es omnicomprensiva, pero proporciona un camino para confrontar el problema fundamental de las categorías antes de recurrir a la psicología social y a todos los problemas que se manifiestan entre la percepción y el lenguaje.

Como cualquier otra teoría científica, la SGN debe ponerse a prueba mediante experimentos. Los datos provenientes de diversas líneas de estudio le proporcionan ya un sustento significativo. Los estudios sobre el desarrollo del sistema nervioso muestran que la regulación dinámica de un pequeño número de moléculas de adhesión de las células es responsable de los patrones de crecimiento de los procesos neuronales,[25] y no hay evidencia del tipo de sistema extensivo de marcadores químicos que sería requerido si toda la conectividad estuviera de algún modo genéticamente especificada. Por cierto, hay evidencia de una varianza extensiva de la conectividad, incluso en animales genéticamente idénticos. En los animales adultos, los datos muestran que los mapas neuronales, como los del córtex somatosensorial (un área del cerebro que responde a receptores de tacto en la piel) no son tan rígidos e invariables como anteriormente se había pensado.[26] Más bien, se da una competencia para determinar los trazados de los mapas. La evidencia muestra que ciertas conexiones que están anatómicamente presentes pero no son usadas en el funcionamiento normal de estas regiones corticales, se pueden activar mediante la selección cuando se perturban los patrones normales de activación. Estas conexiones inactivas corresponden a las que no se seleccionan en el funcionamiento normal; su existencia no tendría sentido en un sistema no selectivo. Una discusión com-

pleta[27] de esta evidencia[28] excede el ámbito de este artículo. En vez de eso, describiremos brevemente una serie de autómatas que se han construido en simulación computada para poner a prueba la consistencia de la SGN, así como para demostrar la habili-dad de los sistemas de reconocimiento selectivos para realizar interesantes tareas de reconocimiento y categorización. Esos modelos pueden ser invalorables para ayudar a que los biólogos se concentren en una serie de cuestiones experimentales. En última instancia, esperamos también aprender de ellos cómo construir máquinas capaces de desarrollar tareas de clasificación sensibles al contexto mucho mejor que cualquier computadora hoy disponible.

Autómatas de reconocimiento selectivo

Hemos estado explorando las propiedades de los sistemas de reconocimiento selectivo mediante la construcción y la puesta a prueba de una serie de autómatas. Estos autómatas abordan algunos de los problemas del paradigma estándar de la IA evitando tanto las categorías preestablecidas como la programación. En lugar de eso, están construidos como redes de unidades similares a las neuronas que, mediante un proceso de selección, pueden realizar tareas simples de categorización y asociación en mundos cambiantes llenos de novedad. La programación se utiliza para instruir a la computadora sobre la forma de emular las unidades neuronales, pero la función de esas unidades no está programada. El primer autómata, llamado Darwin I, afronta el proceso de reconocimiento mismo utilizando hileras de dígitos binarios como cosa a conocer o como reconocedores.[29] El segundo, Darwin II, se utilizó para el reconocimiento y clasificación de patrones bidimensionales presentado en una matriz de tipo retinal.[30] Un tercer sistema, Darwin III, combina las redes de reconocimiento y categorización de Darwin II con circuitos motores y efectores que actúan sobre el ambiente para formar un autómata completo capaz de comportamiento autónomo. A diferencia de los autómatas anteriores, se puede observar que Darwin III se comporta sin que miremos en el interior de su «sistema nervioso».

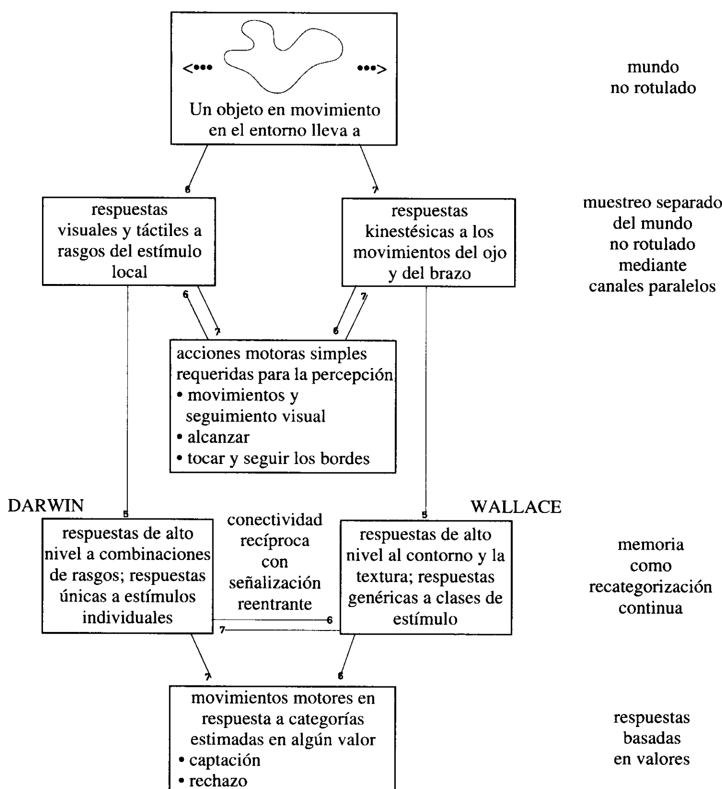
La disposición de las redes en Darwin III puede cambiarse a voluntad para adecuarse a distintos protocolos experimentales. Un esquema funcional simplificado de una de esas disposiciones

—que en seguida describiremos en detalle— puede observarse en el diagrama de la página 189.

Las redes se pueden construir a partir de múltiples repertorios, correspondientes a las regiones funcionales del cerebro. Cada repertorio puede contener varias capas de células, igual que la corteza cerebral. Cada capa puede tener sus propias reglas de conectividad y modificación sináptica. Una vez establecida, la conectividad es fija, pero las fuerzas de conexión varían de acuerdo con reglas para la modificación sináptica que proporcionan los mecanismos de la SGN. Los estímulos se presentan en una matriz de tipo retiniano. Un módulo ambiental permite que esos estímulos se generen y se muevan en diversas formas, para poner a prueba las respuestas del autómata. Se han proporcionado a Darwin III dos dispositivos para el educto motor: un brazo de junturas múltiples y una cabeza móvil con uno o dos ojos. Los movimientos del brazo pueden efectivamente desplazar objetos del entorno, mientras que los movimientos de la cabeza sólo afectan la posición percibida. Se pueden incorporar redes especializadas que responden mediante simples criterios innatos al valor adaptativo que tienen para el autómata sus diversas acciones motoras. Se puede hacer que la amplificación selectiva dependa del valor adaptativo tal como lo registran esas estructuras internas. No se permiten criterios externos para la amplificación, tales como los que proporcionaría un programador.

Una regla importante, que distingue a los autómatas Darwin de los sistemas de IA con objetivos similares, es que no se incorpora al sistema, cuando se lo construye, información específica sobre los objetos de estímulo. La información general sobre las clases de estímulo que serán significativas para el sistema (por ejemplo, el hecho de que habrá dibujos de líneas) está sin embargo implícita en la elección de los elementos detectores de rasgos que serán usados. Esta elección es parecida a las especializaciones inherentes a los órganos receptores de cada especie durante el curso de la evolución.

Una parte de Darwin III está especializada, como Darwin II, para tratar con la categorización. Los estudios psicológicos sugieren que los humanos utilizan diversos métodos para hacer clasificaciones.[31] De acuerdo con ello, ambos autómatas incorporan dos funciones de categorización separadas. La primera es reconocer aspectos de los objetos individuales de acuerdo con sus características únicas; la segunda es reconocer similitudes entre



Este diagrama muestra las relaciones funcionales en una de las disposiciones de Darwin III. La caja superior representa el entorno, en el cual se mueve un objeto sin nombre. Las cajas siguientes representan funciones, cada una de las cuales está asistida por varios repertorios de grupos neuronales con las interconexiones apropiadas. La flecha sugiere relaciones causales, que generalmente reflejan la existencia de conexiones anatómicas entre diversas regiones. Los dos sistemas separados de muestreo que posee el sistema, Darwin y Wallace, están a la izquierda y a la derecha, respectivamente.

El resultado de la actividad neuronal del autómata se hace externamente evidente como actividad motora en respuesta a su categorización de los objetos. Esta categorización procede de acuerdo con criterios internos que emergen porque el autómata tiene sesgos o valores. Por ejemplo, el valor «ver es mejor que no ver» se expresa en términos de cambios en las fuerzas de conexión en los repertorios oculomotores cuando las unidades visuales quedan más activas después de los movimientos del ojo. Nótese que el valor no prescribe categorías, pero cuando las categorías emergen, eso sesga la selección de conductas consecuentes. Nótese también que el autómata refleja su experiencia en alteraciones más o menos estables de las fuerzas de conexión, pero no posee representaciones codificadas para la memoria. En vez de eso, la memoria se exhibe como una habilidad aumentada para reconocer y categorizar los objetos en clases que se han visto antes.

cosas de la misma clase o diferencias entre cosas de distintas clases y definir así los objetos. Para realizar estas tareas, se utilizan dos conjuntos de repertorios operando en paralelo. Ambos incorporan series de mapas, pero los mapas están construidos en forma diferente y hacen clasificaciones de acuerdo con principios muy distintos. Estos subsistemas complementarios interactúan mediante conexiones reentrantes recíprocas que les otorgan funciones asociativas que no posee ningún conjunto por separado.

Se ha llamado arbitrariamente a los dos subsistemas Darwin y Wallace (véase el diagrama). Darwin es el designado para responder únicamente a patrones de estímulo individuales. Sus respuestas responden a grandes trazos a una estrategia de categorización conocida en la literatura psicológica como «apareo con ejemplares». En esta estrategia, los objetos se comparan con ejemplares almacenados, y se asignan categorías de acuerdo con el grado máximo de equivalencias entre los rasgos; Darwin, sin embargo, no almacena patrones de rasgos sino que más bien produce patrones de respuesta correspondientes a cada insumo individual que reconoce. En sí, Darwin no puede definir un objeto, porque es insensible a las continuidades de los rasgos. Wallace, por su parte, está diseñado para responder en la misma forma a diferentes objetos en una clase correlacionando una variedad de rasgos; este proceso a grandes trazos se asemeja a la estrategia de categorización que se conoce como comparación de probabilidades, y no puede en sí misma distinguir entre individuos. Por supuesto, las especificidades de respuesta particulares de Darwin y Wallace (principalmente respuestas ante el contorno en un mundo bidimensional) están previstas para ser meramente ejemplares; otras propiedades de estímulo, tales como el color y la textura, también estarían representadas en los sistemas nerviosos reales.

Las redes Darwin y Wallace poseen ambas estructura jerárquica. Cada una tiene un nivel que se conecta directamente con el insumo sensorial y que trata con rasgos de los estímulos; conectado a él hay un nivel de abstracción o combinación que recibe su insumo principal del primer nivel y que responde a combinaciones de las respuestas elementales que son relevantes para la categorización. El nivel inicial del subsistema de Darwin comprende grupos de células que responden a rasgos locales en la matriz de insumo, tales como segmentos de línea orientados en ciertas direcciones o con ciertos dobleces. Estos grupos están conectados

en diversas combinaciones a grupos de abstracción de más alto nivel que dan las respuestas únicas deseadas a cada patrón de estímulo. Estas respuestas incluyen elementos aportados por el contexto circundante.

Por otra parte, Wallace trata con objetos y propiedades de clases. En Darwin III, el subsistema Wallace hace uso del brazo del autómata para seguir el rastro del contorno de los objetos de estímulo, en forma muy parecida a la de un ciego que lee un texto en braille. El repertorio de primer nivel de Wallace recibe insumos de las neuronas sensoriales kinestésicas del brazo. Responde a correlaciones de las actividades de rastreo que distinguen a los objetos como entidades que se destacan del fondo por su continuidad espacial. Las células de este repertorio están conectadas a su vez a una red abstracta, similar a la de Darwin. Dado que el rastreo responde a la presencia o cambio de dirección de las líneas del ambiente, manifestando poco interés por sus longitudes u orientaciones, Wallace es insensible tanto a las transformaciones rígidas como a las no rígidas de los objetos de estímulo y tiende a responder a las características de clases de familias enteras de objetos relacionados.

Las redes en los dos repertorios de abstracción se hallan conectadas entre sí a niveles más altos para formar una pareja de clasificación. Es importante que en esa pareja ambos niveles no estén contectados, porque si así fuera, se confundirían los modos separados de muestreo para la clasificación ulterior y se perderían sus características distintivas. En otras palabras, el muestreo del mundo por canales separados debe ser disyuntivo para que se manifieste una clasificación rica y sensible al contexto.

Las respuestas de los grupos neuronales en Darwin III están determinadas por sus insumos presentes y sus historias pasadas, de una manera que incorpora los rasgos más importantes de las respuestas mucho más complicadas propias de las neuronas reales. Las conexiones de insumo están especificadas por listas que se construyen cuando se configura el modelo. Estas listas se construyen de modo diferente para cada repertorio, de acuerdo con su función. Cada insumo contribuye al educto del grupo de acuerdo con la fuerza actual de su conexión. Además, los grupos están sujetos a fluctuaciones al azar en su actividad, análogas a las que se encuentran en las redes neuronales reales. La especificidad de reconocimiento de cada grupo depende de sus listas de

conexiones y de sus fuerzas de conexión; la mejor respuesta se obtiene cuando los insumos más activos se conectan vía sinapsis con altas fuerzas de conexión. Para los otros insumos, un grupo responderá más o menos bien, superponiéndose en especificidad a otros grupos y confiriendo degeneración al sistema como un todo.

Las fuerzas de conexión entre células, dentro y fuera de los grupos, se modifican durante la amplificación selectiva. La regla de amplificación depende sólo de las cantidades que pueden tener influencia en la eficacia de las sinapsis reales en las redes neuronales verdaderas y es puramente de naturaleza local. En el esquema que hemos usado más a menudo, una conexión se refuerza si las actividades de las células pre y post-sinápticas exceden los umbrales especificados. En otras palabras, una conexión que va desde un insumo activo a una célula activa se refuerza y conduce a una respuesta más fuerte la próxima vez que se encuentra un insumo similar. En ciertas otras circunstancias, las fuerzas de conexión se debilitan, impidiendo a veces que el sistema alcance un estado en el que todas las sinapsis tengan la máxima eficacia. Si esto llegara a pasar, cualquier insumo conduciría a la red como un todo a un estado de máxima actividad correspondiente a una especie de «acceso epiléptico».

Con secuencias de estimulación apropiadas, Darwin II y Darwin III son capaces de producir respuestas correspondientes a conductas tales como la categorización, el reconocimiento, la generalización y la asociación. En otra parte hemos presentado un tratamiento detallado de algunos experimentos típicos.[32] La *categorización* es más evidente en las respuestas de Wallace. Esta no involucra denominación, lo que requiere una convención lingüística, sino sólo similitudes de respuesta a ítem de la misma categoría. Como ya hemos visto, esa similitud es característica de la conducta de Wallace. (Las categorías particulares a las que se llega para diversas especies de estímulo dependen de la elección particular de correlaciones de rasgos kinestésicos a las que responden los grupos de Wallace, y pueden o no estar de acuerdo con las categorías que *nosotros* definimos para el estímulo. La existencia de discrepancia es consistente con la idea de que las categorías no son inherentes al entorno, sino que dependen de las disposiciones evolutivamente dictadas del organismo para prestar atención a las categorías que son relevantes a sus necesidades

adaptativas. La construcción de información depende de esos criterios adaptativos internos.)

El efecto de la modificación sináptica sobre las respuestas de los grupos en estas redes es el de alterar las células que responden por encima de un cierto nivel (el umbral de amplificación), de modo que den una respuesta más fuerte cuando más tarde experimenten un patrón de estímulo similar; los grupos con respuestas más débiles en general se cambian como para no responder en absoluto después de la amplificación. Los grupos que no están involucrados en la respuesta a un estímulo particular permanecen sin cambios y disponibles para responder a otros estímulos que aún no se han encontrado. Estos cambios selectivos demuestran *reconocimiento*, o el incremento de la respuesta significativa a un estímulo que ya se ha experimentado antes. Por añadidura, se puede demostrar que la amplificación aumenta la habilidad del sistema para categorizar.

La *generalización* ocurre cuando la respuesta a las nuevas formas se parece más a las respuestas de formas encontradas antes en la misma clase de lo que sería el caso sin una experiencia anterior. En Wallace, la generalización ya está presente sin amplificación como consecuencia de la propiedad de correlación de rasgos de esa red, pero en Darwin la generalización no está preincorporada y sólo puede ocurrir con la ayuda de Wallace. Las conexiones reentrantes entre Darwin y Wallace hacen posible que los patrones de respuesta en Wallace influyan en las unidades amplificadas en Darwin de acuerdo con la pertenencia a clases de los diversos estímulos. Tras un número de repeticiones de este proceso, con contornos de varias clases diferentes, las respuestas de Darwin llegan a ser más parecidas dentro de cada clase. En la medida en que los estímulos nuevos de la misma clase eliciten respuestas en la capa detectora de rasgos de Darwin que tengan elementos en común con las respuestas a los estímulos utilizadas durante la amplificación, las respuestas a estos nuevos estímulos también llegarán a ser más parecidas, consistentemente con la generalización en Darwin.

Las conexiones reentrantes en ambas direcciones entre Darwin y Wallace también son esenciales para los procesos de *asociación*. La asociación se logra cuando las diversas respuestas en Darwin se vinculan a diferentes estímulos en la misma clase a través de Wallace, de tal manera que la presentación de uno de los

estímulos individuales evoque elementos de la respuesta en Darwin que es apropiada para otro. Wallace no puede exhibir separadamente esas propiedades asociativas entre individuos.

Estas capacidades de categorización, reconocimiento, generalización y asociación subrayan el hecho de que esos aspectos críticos de la percepción pueden e incluso deben ocurrir ante el aprendizaje convencional. También muestran la forma en que los sistemas basados en más de un principio de categorización pueden unirse en parejas de clasificación para dar modos clasificatorios no disponibles en ningún sistema por separado.

También se usa a Darwin III para estudiar problemas que involucran movimiento, constante perceptual, discriminación entre figura y fondo y memoria, entre otros fenómenos. La teoría de la SGN sugiere que el movimiento del objeto es un factor crítico en el proceso selectivo, particularmente en el aprendizaje visual temprano, donde proporciona al sistema perceptual el principal indicio de que el mundo puede ser de hecho separado en objetos distintos. Esta perspectiva es consistente con experimentos que sugieren que los bebés humanos poseen una concepción de los objetos como entidades espacialmente conectadas, coherentes, continuas y móviles.[33] Por esta razón, los primeros experimentos que utilizan los sistemas motores en Darwin III han sido diseñados para ganar una mejor comprensión de este proceso de análisis, comenzando con la simple habilidad para seguir el rastro de un objeto móvil sobre un fondo plano.

En estos experimentos, las capas visuales mapeadas retinotípicamente (correspondientes al colículo superior o tecto óptico en diversas especies biológicas) proporcionan conexiones a las capas motoras conectadas en última instancia a los músculos que controlan la posición del ojo. Estas conexiones ligan indiscriminadamente células motoras correspondientes a todas las direcciones del movimiento ocular. De este modo, no hay una especificidad a priori con respecto al insumo de una región particular del campo visual y no hay habilidades motoras dispuestas de antemano; la performance sólo puede desarrollarse y perfeccionarse mediante la selección, a partir de movimientos espontáneos generados por pares de capas motoras generadoras de patrones mutuamente inhibitorias. La amplificación de conexiones de la región sensorial a la motora se halla modulada por esquemas de valor que se basan respectivamente en la apariencia de la activi-

dad en una región circunfoveal y en la fóvea misma. (La fóvea es la región de mayor agudeza visual, cerca del centro de la retina.) De esta manera, las conexiones de un punto particular en la capa visual con un área motora particular tienden a reforzarse cuando la apariencia de un objeto en ese punto se correlaciona con la actividad en el área motora que lleva a la foveación del estímulo. La misma conexión tenderá a debilitarse cuando la actividad motora no conduzca a la foveación.

Después de un período de experiencia adecuado con diversos estímulos móviles, Darwin III de hecho comienza a realizar los movimientos de rastreo más finos y apropiados sin más especificación ulterior de su tarea que la que se halla implícita en el esquema de valor. El autómata finalmente exhibe un sistema de conducta en el cual el ojo barre al azar cuando no hay ningún estímulo visible, hace un movimiento rápido ante cualquier estímulo que aparece dentro de los límites de su campo visual más amplio y finalmente rastrea cualquier estímulo que ha sido exitosamente foveado. Durante el seguimiento fino, las redes Darwin y Wallace son capaces de responder al objeto ahora centrado, permitiendo que ocurra una categorización independiente de la posición. Eventualmente se configura un hábito, permitiendo ocasionales movimientos a otras partes del campo visual. Después de uno de esos movimientos, se toma otro nuevo objeto de estímulo como blanco de rastreo.

En forma parecida, el brazo de junturas múltiples de Darwin III se puede entrenar para alcanzar y tocar los objetos que antes detecta y sigue el sistema visual. Esta performance, que entraña la coordinación de movimientos gestuales que comprometen a diversas junturas en diferentes medidas, requiere la participación de una amplia serie de repertorios que ejecutan funciones similares a las que desarrolla el cerebelo en el sistema nervioso.

Por supuesto, estos sistemas selectivos cometen errores. Por ejemplo, en el sistema oculomotor pueden manifestarse errores de amplificación cuando un estímulo cruza diagonalmente el borde de un campo visual, y estímulos muy grandes pueden confundir al mecanismo de seguimiento. Mediante el uso de técnicas de ingeniería estándar, se podría haber diseñado sin duda un mejor sistema de seguimiento, uno en el que se pudieran calcular por anticipado e incorporar a la lógica del diseño todos los movimientos correctos para una zona de luz en cualquier ubicación del

campo visual. El sistema selectivo necesita experiencia para desarrollar sus capacidades funcionales. Pero esta misma falta de funciones preconstruidas es su mayor ventaja; la maquinaria en un sistema selectivo «no sabe para qué es», permitiendo que las mismas redes cumplan diversas tareas dependiendo de lo que nosotros, actuando como agentes externos de la «evolución», decidimos que ellas encontrarán «adaptativo». Esta flexibilidad es posible, precisamente, porque la organización funcional del sistema sólo surge después de la interacción con el ambiente. Regímenes de entrenamiento igualmente simples funcionarían para una amplia variedad de tareas, tales como encontrar un objeto deseado sobre un fondo de objetos que distraen y tomarlo con el brazo. La combinación modular de subsistemas capaces de desarrollar tareas de reconocimiento proporciona una estrategia atractiva para la construcción de autómatas con capacidades perceptuales de creciente complejidad, y tendrá a su tiempo significación en los esfuerzos en pos de la visión de máquina. Pues, como producto, el corolario del control motor es la habilidad para generar secuencias de comportamiento complejas mediante la selección a partir de patrones de movimiento simples e innatos. Esto posee una obvia significación para el desarrollo de la robótica y del control de sistemas complejos, particularmente en conexión con «tareas atencionales».[34] Tales tareas sensibles al contexto presentan una barrera aparentemente insalvable a las estrategias tradicionales de IA.

Desde un punto de vista puramente económico, puede argumentarse que los sistemas selectivos artificiales serían imprácticos porque siempre deben contener algunas unidades que no responden a algunos de los insumos que efectivamente se presentan, y por lo tanto no los utilizan en absoluto. La fracción de estas unidades no aumenta con el tamaño del sistema, sin embargo, y las unidades extras proporcionan una redundancia que puede repercutir en ahorros generales de costo, dadas las pocas reparaciones y el escaso tiempo de paro. La velocidad de ejecución tampoco es una barrera para la utilidad de los sistemas selectivos implementados en computadoras en paralelo; la respuesta sólo requiere unas pocas veces el tiempo de latencia de las unidades individuales, porque no se requiere relajamiento, como en los sistemas basados en el concepto mecánico-estadístico de minimización de la energía de la red. Es por lo tanto razonable esperar que

se construyan sistemas prácticos basados en la selección con hardware actualmente disponible o previsible en un futuro cercano. Estos sistemas tendrán más componentes, pero un sistema global más simple que los posibles sistemas basados en modelos de red de procesamiento de información. Obviamente, tales sistemas, si se prueban viables, tendrán una enorme influencia en el diseño del hardware de computación.

Resumen y conclusiones

Los autómatas que hemos descripto intentan ilustrar ciertos aspectos de la teoría de la SGN sin pretender emular al sistema nervioso verdadero en forma detallada. Los datos experimentales y los recursos de computación para esto último no están todavía a la mano. De este modo, los modelos no proporcionan evidencia para la aplicabilidad de la teoría a sistemas nerviosos reales; esa evidencia sólo puede venir de la experimentación. Sin embargo, la SGN proporciona explicaciones satisfactorias para una panoplia de procesos perceptuales que involucran categorización y, como ejemplos inanimados de sistemas selectivos, los modelos pueden ayudar a demostrar la consistencia de la teoría como descripción abstracta de esos procesos. Estas demostraciones son importantes para comprender los complicados sistemas biológicos; al mismo tiempo pueden proporcionar una comprensión real del problema de la computación científica de diseñar sistemas artificiales con capacidades similares a las del cerebro.

En particular, está muy claro que los sistemas nerviosos no trabajan en nada que se parezca a la forma que se ha supuesto en el paradigma estándar de la IA, aunque la performance de al menos algunos sistemas nerviosos sea definitoria del término *inteligencia* tal como se usa en la frase «inteligencia artificial». Es entonces muy curioso que la IA, aun en su nueva guisa conexionista, haya descuidado en su mayor parte la biología fundamental del sistema nervioso, del cual la noción misma de inteligencia se deriva. Sugerimos que para progresar en la superación de los obstáculos que hemos discutido, la IA debe reconocer esos orígenes e incorporar lo que hemos aprendido del estudio de los sistemas nerviosos. Debe dejar de razonar por analogía con sistemas físicos bien estudiados, pero irrelevantes, como los cristales

de spin, y en vez de eso debe razonar analizando los hechos relevantes sobre los sistemas biológicos que realmente tienen inteligencia. Este cambio exigirá que la IA abandone la noción de inteligencia como una actividad puramente abstracta de procesamiento de la información.

Una confrontación del número de niveles interactivos en un organismo real capaz de conducta inteligente revela una estremecedora complejidad de interacciones no lineales. Si se agrega la transmisión social a través del lenguaje, la complejidad se incrementa aún más. A la vista de esta complejidad, parece ser el colmo de la arrogancia pensar que todos los problemas que confrontan las criaturas inteligentes se pueden comprender ponderándolos en abstracto. En vez de eso, se debe comenzar a analizar esos sistemas en términos de las estructuras y funciones básicas necesarias y sus modos de origen, su desarrollo tanto como su evolución. La separación del hardware y el software implícita en la estrategia de la IA tradicional tiene que abandonarse, aun cuando haya servido como principio orientador en el desarrollo de las computadoras de tipo von Neumann, que pueden ser máquinas lógicas —y en alguna medida, máquinas culturales— pero que no son máquinas biológicas. Creemos que la IA finalmente sólo se alcanzará en sistemas no-von Neumann en los que variantes especializadas de hardware, basadas en el tema común de la selección y el pensamiento de la población, trabajarán sin programas para adaptarse a los ambientes particulares en los que se encuentren, tal como lo hacen los organismos biológicos. Los programas y la inteligencia basada en la comunicación podrán venir después.

Notas

Agradecemos a la International Business Machines Corporation y a la Cornell National Supercomputer Facility por su ayuda en parte de este trabajo.

[1] Aristóteles, *Historia Animalium*, traducción de D'Arcy Wentworth Thompson (Oxford: Clarendon Press, 1910), libro 2, sección 3, 501^b 20.

[2] Aristóteles, *De Partibus Animalium*, traducción de William Ogle (Oxford: Clarendon Press, 1912), libro 3, sección 1, 661^b 33.

[3] Gerald M. Edelman, *Neural Darwinism: The Theory of Neuronal Group Selection*(Nueva York: Basic Books, 1987).

[4] Patrick Henry Winston, *Artificial Intelligence*, 2ª edición (Reading, Mass.: Addison-Wesley, 1984), 1-2.

[5] Marvin Minsky, «Steps Toward Artificial Intelligence», en *Proceedings of the Institute of Radio Engineers*, 49 (1961):8-30.

[6] Winston, *Artificial Intelligence*, 42.

[7] Hubert L. Dreyfus, *What Computers Can't Do* (Nueva York: Harper and Row, 1972).

[8] Alan M. Turing, «On Computable Numbers, with an Application to the Entscheidungsproblem», en *Proceedings of the London Mathematical Society* 42 (1937):230-265. Turing definió una clase determinada de autómata, conocida hoy como la máquina de Turing, y demostró que cualquier miembro de esta clase podía computar cualquier función entre un gran conjunto de clases. Todas las computadoras digitales, excepto unas pocas de propósitos especiales, son máquinas de Turing.

[9] Douglas R. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid* (Nueva York, Basic Books, 1979), 561 [Traducción española: *Gödel, Escher, Bach. Un eterno y grácil bucle*, Barcelona, Tusquets, 1987]. Hofstadter señala que «como el té, la tesis de Church-Turing se puede dar en toda una variedad de diferentes fuerzas», y presenta varias de ellas con una discusión de sus implicaciones para la IA.

[10] John McCarthy, «Programs with Common Sense», en *Proceedings of the Teddington Conference on the Mechanization of Thought Processes* (Londres: Her Majesty's Stationery Office, 1960); John McCarthy, «Some Expert Systems Need Common Sense», *Annals of the New York Academy of Sciences* 426 (1984):129-35.

[11] Peter W. Frey, editor, *Chess Skill in Man and Machine*, 2ª edición (Nueva York: Springer-Verlag, 1984).

[12] Raymond Reiter, «Nonmonotonic Reasoning», *Annual Reviews of Computer Science* (Palo Alto: Annual Reviews Inc de próxima aparición).

[13] Warren S. McCulloch y Walter H. Pitts, «A Logical Calculus of the Ideas Immanent in Nervous Activity», *Bulletin of Mathematical Biophysics* 5 (1943):115-33.

[14] Frank Rosenblatt, *The Perceptron: A Theory of Statistical Separability in Cognitive Systems* (Ithaca: Cornell Aeronautical Laboratory Inc., Report Nº VG-1196-G-1, 1958).

[15] Marvin Minsky y Seymour Papert, *Perceptrons: An Introduction to Computational Geometry* (Cambridge: MIT Press, 1969).

[16] David Marr y Tomaso Poggio, «Cooperative Computation of Stereo Disparity», *Science* 194 (15 de octubre de 1976):283-87.

[17] John R. Anderson, «A Spreading Activation Theory of Memory», *Journal of Verbal Learning and Verbal Behavior* 22 (1983):261-95.

[18] Leon N. Cooper, «A Possible Organization of Animal Memory and Learning», en *Nobel Symposium*, nº 24 (1973), 252-64; Teuvo Kohonen, *Associative Memory: A System Theoretical Approach* (Nueva York: Springer-Verlag, 1977).

[19] Revisadas en John J. Hopfield y David W. Tank, «Computing with Neural Circuits: A Model», *Science* 233 (8 de agosto de 1986):625-33.

[20] David H. Ackley, Geoffrey E. Hinton y Terrence J. Sejnowski, «A Learning Algorithm for Boltzmann Machines», *Cognitive Science* 9 (enero-marzo de 1985):147-69.

[21] David E. Rumelhart, James L. McClelland y el PDP Research Group, editores, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volúmenes 1 y 2 (Cambridge: MIT Press, 1986).

[22] Edelman, *Neural Darwinism*.

[23] Ibid.; y Gerald M. Edelman, «Group Selection and Phasic Reentrant Signalling: A Theory of Higher Brain Function», en *The Mindful Brain*, Gerald M. Edelman y Vernon B. Mountcastle, eds. (Cambridge: MIT Press, 1978), 51-100.

[24] Edelman, *Neural Darwinism*.

[25] Gerald M. Edelman, «Cell-Adhesion Molecules: A Molecular Basis for Animal Form», *Scientific American* 250 (abril de 1984): 118-29; Gerald M. Edelman, «Cell Adhesion Molecules in the Regulation of Animal Form and Tissue Pattern», *Annual Review of Cell Biology* 2 (1986): 81-116.

[26] John H. Kaas, Michael M. Merzenich y Herbert P. Killackey, «The Reorganization of Somatosensory Cortex Following Peripheral-Nerve Damage in Adult and Developing Mammals», *Annual Review of Neuroscience* 6 (1983): 325-56; Michael M. Merzenich et al., «Topographic Reorganization of Somatosensory Cortical Areas 3b and 1 in Adult Monkeys Following Restricted Deafferentation», *Neuroscience* 8 (enero de 1983):33-55. Se han reportado muchos otros ejemplos.

[27] Edelman, *Neural Darwinism*.

[28] Gerald M. Edelman y Leif H. Finkel, «Neuronal Group Selection in the Cerebral Cortex», en *Dynamical Aspects of Neocortical Function*, Gerald M. Edelman, W. Einer Gall y W. Maxwell Cowan, eds. (Nueva York: John Wiley, 1984):653-95.

[29] Gerald M. Edelman, «Group Selection as the Basis for Higher Brain Function», en *Organization of the Cerebral Cortex*, Francis O. Schmitt et al., eds. (Cambridge: MIT Press, 1981), 51-100.

[30] Gerald M. Edelman y George N. Reeke (h) «Selective Networks Capable of Representative Transformations, Limited Generalizations, and Associative Memory», en *Proceedings of the National Academy of Sciences USA* 79 (1982):2091-95.

[31] Edward E. Smith y Douglas L. Medin, *Categories and Concepts* (Cambridge: Harvard University Press, 1981).

[32] George N. Reeke, (h) y Gerald M. Edelman, «Selective Networks and Recognition Automata», *Annals of the New York Academy of Sciences* 426 (1984):181-201.

[33] Philip J. Kellman y Elizabeth S. Spelke, «Perception of Partly Occluded Objects in Infancy», *Cognitive Psychology* 15 (4) (octubre de 1983):483-524.

[34] Anya Hurlbert y Tomaso Poggio, «Do Computers Need Attention?», *Nature* 321 (12 de junio de 1986): 651-52.

8
La inteligencia como conducta emergente, o la canción del Edén

W. Daniel Hillis

A veces un sistema con muchos componentes simples exhibirá una conducta global que parece más organizada que la conducta de las partes individuales. Consideremos la estructura intrincada de un copo de nieve. Las formas simétricas en los cristales se repiten en tres y seises, con patrones recurrentes de un lado a otro y dentro de sí mismos a diferentes escalas. Las formas asumidas por el hielo son consecuencias de reglas locales de interacción que gobiernan las moléculas de agua, aunque las conexiones entre las formas y las reglas están lejos de ser obvias. Después de todo, éstas son las mismas tres reglas de interacción que hacen que el agua se transforme súbitamente en vapor en su punto de hervor y cause remolinos que forman una turbulencia. Las reglas que gobiernan las fuerzas entre las moléculas de agua parecen mucho más simples que los cristales o los remolinos o los puntos de hervor, aunque todos estos fenómenos complejos son de algún modo consecuencia de esas reglas. Esos fenómenos se denominan *conductas emergentes* del sistema.

Sería muy conveniente que la inteligencia fuera una conducta emergente de neuronas conectadas al azar, en el mismo sentido que los copos de nieve y los remolinos son conductas emergentes de las moléculas de agua. Sería posible entonces construir una máquina pensante simplemente enganchando una cantidad suficientemente grande de neuronas artificiales. La noción de emergencia sugiere que una red tal, una vez que ha alcanzado una masa crítica, espontáneamente comenzaría a pensar.

W. Daniel Hillis. Diseñador de la «Máquina de Conexión», una computadora paralela que fue el tema de su tesis de doctorado en el Instituto Tecnológico de Massachusetts.

Esta es una idea seductora porque abre la posibilidad de construir inteligencia sin tener que comprenderla primero. La comprensión de la inteligencia es difícil y probablemente esté muy lejos de lograrse, de modo que la posibilidad de que ella pueda emerger espontáneamente de las interacciones de una gran colección de partes simples posee un considerable atractivo para los potenciales constructores de máquinas pensantes. Desafortunadamente, la idea no sugiere una aproximación práctica a esa construcción. El concepto de emergencia en sí mismo no ofrece ni una guía para construir ese sistema ni una comprensión de la forma en que podría trabajar.

Irónicamente, la inescrutabilidad aparente de la idea de inteligencia como conducta emergente da cuenta de gran parte de su creciente popularidad. La emergencia ofrece una forma de pensar en la causalidad física, manteniendo simultáneamente la imposibilidad de una explicación reduccionista del pensamiento. Para quienes temen las explicaciones reduccionistas de la mente humana, nuestra ignorancia de la forma en que las interacciones locales producen conducta emergente ofrece una niebla de reaseguro en la cual se puede esconder el alma.

Recientemente se ha renovado el interés en la conducta emergente en forma de redes neuronales simuladas y modelos conexionistas, cristales de spin, autómatas celulares y modelos evolucionistas. Cada uno de ellos es el modelo de algún sistema real. Para las redes neuronales y los modelos conexionistas, el sistema que se está modelando es una colección de neuronas biológicas, tal como el cerebro; para los cristales de spin es un cristal molecular. Los autómatas celulares y los modelos evolucionistas se basan en la ontogénesis y filogénesis de organismos vivientes. En todos estos casos, tanto el modelo como el sistema que se está modelando producen dramáticos ejemplos de conducta emergente.

La mayoría de estos modelos no es nueva, pero el interés en ellos se está agitando merced a una combinación de nuevas comprensiones y nuevas herramientas. Las comprensiones se originan primariamente en una rama de la física que se llama teoría de los sistemas dinámicos. Las herramientas provienen del desarrollo de nuevos tipos de dispositivos computacionales. Así como en la década de 1950 pensábamos la inteligencia en términos de servomecanismos, y en los sesenta y setenta en términos de computadoras secuenciales, ahora estamos comen-

zando a pensarla en términos de computadoras paralelas, en las que decenas de miles de procesadores trabajan juntos. Este no es un giro filosófico profundo, pero es de gran importancia práctica, dado que ahora es posible estudiar grandes sistemas emergentes en forma experimental.

Inevitablemente, los antirreduccionistas interpretan esos progresos como un cisma entre los racionalistas simbólicos, quienes les son opuestos, y los gestálticos, quienes les dan sustento. A menudo me han preguntado de qué lado estoy. No siendo un filósofo, me inclino a concentrarme en los aspectos prácticos de la pregunta: ¿Cómo podríamos comenzar a construir una inteligencia emergente? ¿Qué información necesitaríamos conocer para tener éxito? ¿Cómo se puede determinar experimentalmente esa información?

El sistema emergente que puedo imaginar con más facilidad sería una implementación del pensamiento simbólico antes que una refutación de él. El pensamiento simbólico sería una propiedad emergente del sistema. El punto de vista se explica mejor por la siguiente parábola sobre el origen de la inteligencia humana. Por lo que yo sé, esta parábola de la evolución humana es consistente con la evidencia disponible (como lo son muchas otras), pero dado que se ha escogido para ilustrar el punto debe leerse como una historia, no como una teoría. Es diferente de la mayoría de las teorías aceptadas sobre el desarrollo humano en que ella presenta rasgos que pueden medirse en los registros arqueológicos —tales como volumen creciente del cerebro, distribución de alimentos y neotenia— como consecuencias más que como causas de la inteligencia.

Había una vez, hace cerca de dos millones y medio de años, una raza de simios que caminaba erguida. En términos de intelecto y de hábitos eran parecidos a los modernos chimpancés. Los simios jóvenes, como los simios jóvenes de hoy en día, tenían tendencia a imitar las acciones de los demás. En particular, tenían tendencia a imitar sonidos. Si un simio chillaba «ooh, ehh, ehh», otro repetiría «ooh eeh eeh». Algunas secuencias de sonidos, o «canciones», se repetían con más probabilidad que otras.

Ignoremos la evolución de los simios por el momento y consideremos la evolución de las canciones. Dado que las canciones eran replicadas por los simios, y dado que a veces morían y en ocasiones se combinaban con otras, podemos considerarlas, a muy grandes trazos, como formas de vida. Sobrevivían, se mezcla-

ban, competían entre sí y evolucionaban conforme a su propio criterio de adecuación. Si una canción contenía una frase particularmente pegadiza que hacía que se la repitiera a menudo, entonces esa frase tenía probabilidad de que se la repitiera en otras canciones. Sólo sobrevivían las canciones que tenían una fuerte tendencia a ser repetidas.

La supervivencia de una canción sólo estaba relacionada indirectamente con la supervivencia de los simios; se hallaba más directamente afectada por la supervivencia de otras canciones. Dado que los simios eran un recurso limitado, las canciones tenían que competir entre sí para tener chance de ser cantadas. Para una canción, una estrategia de competencia exitosa era la de especializarse esto es, encontrar un nicho especial en el cual fuera apta para ser repetida. Las canciones que se adecuaban particularmente bien con estados de ánimo específicos o actividades específicas de los simios tenían por esta razón un especial valor de supervivencia. (No sé por qué algunas canciones encajan bien con ciertos estados de ánimo, pero dado que esto es verdad para mí, no me parece forzado creer que también lo fue para mis ancestros.)

Antes que las canciones comenzaran a especializarse no eran particularmente de valor para los simios. En un sentido biológico, las canciones eran parásitos que tomaban ventaja de la tendencia de los simios a imitar. A medida que las canciones se especializaron, sin embargo, se tornó ventajoso para los simios prestar atención a las canciones ajenas y diferenciarlas. Escuchando las canciones, un simio sagaz podía obtener información útil. Por ejemplo, un simio podía inferir que otro simio había hallado comida o que era probable que atacara. Una vez que los simios comenzaron a sacar ventaja de las canciones, se desarrolló una relación simbiótica: las canciones incrementaban su propia supervivencia proporcionando a los simios información útil; los simios incrementaban la suya mejorando su capacidad para recordar, reproducir y comprender las canciones. De este modo las fuerzas ciegas de la evolución crearon una sociedad entre las canciones y los simios que prosperó sobre las bases del mutuo interés. Con el tiempo esta sociedad evolucionó, transformándose en uno de los simbiontes más exitosos del mundo: la raza humana.

Por desdicha las canciones no dejan fósiles, de modo que a menos que algún proceso natural haya dejado algún rastro fonográfico, nunca sabremos si el relato precedente describe lo

que realmente sucedió. Pero si la historia es verdadera, los simios y las canciones se convirtieron en los dos componentes de la inteligencia humana. Las canciones evolucionaron en conocimiento, moral y mecanismos de pensamiento que en su conjunto constituyen la porción simbólica de la inteligencia humana. Los simios se convirtieron en simios con cerebros más grandes, quizás optimizados para una madurez prolongada, a fin de que pudieran aprender más canciones. El *Homo sapiens* es una combinación cooperativa de los dos.

No es insólito en la naturaleza que dos especies vivan juntas con tanta interdependencia que parecerían ser un solo organismo. Los líquenes son simbiontes de un hongo y un alga que viven tan estrechamente entrelazados que sólo se los puede separar bajo el microscopio. Las plantas de habichuelas necesitan que bacterias que viven en sus raíces fijen el nitrógeno que ellas extraen del suelo, y en retorno las bacterias necesitan nutrientes de las plantas de habichuelas. Incluso el *Paramecium bursarra* unicelular utiliza algas verdes que viven dentro de él para sintetizar el alimento.

Otro ejemplo de dos formas de «vida» completamente diferentes que forman una simbiosis puede estar aún más cerca del ejemplo de los simios y las canciones. En *The Origins of Life*, Freeman Dyson sugiere que la vida biológica es una combinación simbiótica de dos entidades autorreproductivas diferentes con formas muy diferentes de replicación.[1] Dyson sugiere que la vida se originó en dos etapas. Mientras que la mayoría de las teorías sobre los orígenes de la vida comienza con nucleótidos que se replicaban en algún «caldo primordial», la teoría de Dyson comienza con gotas de aceite que se metabolizan.

En el comienzo, estas hipotéticas gotas de aceite replicantes no poseían material genético, sino que eran sistemas químicos autoperpetuantes que absorbían materias primas de sus entornos. Cuando una gota alcanzaba cierto tamaño se fragmentaba en dos; más o menos la mitad de sus constituyentes iba para cada parte. Esas gotas desarrollaron sistemas metabólicos eficientes, aun cuando sus reglas de replicación eran muy distintas de las reglas mendelianas de la vida moderna. Una vez que las gotas de aceite se hicieron buenas metabolizadoras, fueron infectadas por otra forma de replicadores que, como las canciones, no poseían metabolismo propio. Eran moléculas parásitas de ADN; como los virus modernos, sacaron ventaja de la maquinaria existente en las

células anfitrionas para reproducirse. Las metabolizadoras y el ADN evolucionaron con el tiempo en la simbiosis de mutuo beneficio que hoy conocemos como vida.

Esta teoría de la vida en dos partes no está muy lejos conceptualmente de la historia en dos partes de la inteligencia. Ambas sugieren que un mecanismo homeostático preexistente fue infectado por un parásito oportunista. Las dos partes se reprodujeron de acuerdo con diferentes conjuntos de reglas, pero supieron coevolucionar tan exitosamente que el simbionte resultante parece ser una sola entidad. Vista a esta luz, la elección entre la emergencia y la computación simbólica en el estudio de la inteligencia es como la elección entre el metabolismo y la replicación genética en el estudio de la vida. Así como el sistema metabólico proporciona un sustrato en el que puede trabajar el sistema genético, así un sistema emergente proporciona el sustrato en el que puede operar el sistema simbólico.

En realidad, el sistema metabólico de la vida es demasiado complejo como para que lo comprendamos plenamente o lo reproduzcamos. En comparación, las reglas mendelianas de replicación genética son casi triviales, y es posible estudiarlas como un sistema en sí mismo sin preocuparnos por los detalles del metabolismo que las sustenta. En el mismo sentido, parece probable que el pensamiento simbólico se pueda estudiar fructíferamente y quizá recrear sin preocuparnos por los detalles del sistema emergente que lo sustenta. Hasta ahora ésta ha sido la estrategia dominante en inteligencia artificial, y también la estrategia que ha experimentado el mayor progreso.

La otra estrategia consiste en construir un modelo del sustrato emergente de la inteligencia. Este sustrato artificial del pensamiento no necesitaría imitar en detalle los mecanismos del sistema biológico; pero sería necesario que exhibiera las propiedades emergentes que se necesitan para sustentar las operaciones del pensamiento.

¿Qué es lo mínimo que necesitamos comprender para construir un sistema así? Por un lado, necesitamos saber cómo será de grande el sistema a construir. La teoría de la información sugiere que la unidad de medida apropiada es el número de dígitos binarios o bits que se requieren para almacenar la información. ¿Cuántos bits se requieren para almacenar la porción adquirida de conocimiento humano de un humano típico? Necesitamos tener una respuesta aproximada para construir una inteligencia emergente con un comportamiento parecido al humano. En la actualidad, la cantidad de información adquirida almacenada por

un cerebro humano promedio no es conocida ni siquiera dentro de los dos órdenes de magnitud, pero en principio se la puede determinar experimentalmente. Hay al menos tres formas para estimar los requerimientos de magnitud de la inteligencia emergente.

Una forma podría ser a través de la comprensión de los mecanismos físicos de la memoria en el cerebro humano. Si la información se almacena primariamente mediante la modificación de sinapsis, entonces sería posible medir la capacidad de almacenamiento de información del cerebro contando el número de sinapsis. En otra parte en esta edición de *Dædalus*, Jacob T. Schwartz estima que el cerebro contiene alrededor de 10^{15} sinapsis. Cada sinapsis puede almacenar varios bits. Pero aun conociendo la cantidad exacta de almacenamiento físico en el cerebro, no quedaría completamente respondida la pregunta sobre el almacenamiento requerido, pues gran parte de la capacidad potencial de almacenamiento puede ser no usada o usada en forma ineficiente. Pero al menos este método puede ayudarnos a establecer un límite superior para esos requerimientos.

Un segundo método para estimar los requerimientos de almacenamiento de la inteligencia emergente consiste en medir la información del conocimiento simbólico mediante alguna forma de muestreo estadístico. Por ejemplo, es posible estimar el tamaño del vocabulario de un individuo poniéndolo a prueba con palabras muestreadas al azar de un diccionario. La fracción de palabras de prueba conocida por el individuo es un buen indicador de la fracción de palabras que él o ella conocen del diccionario completo. El tamaño estimado del vocabulario es la fracción de prueba multiplicada por el número de palabras del diccionario. Ese experimento depende de que se disponga de un cuerpo predeterminado de conocimiento contra el cual medir. Por ejemplo, sería posible estimar cuántos hechos de la *Enciclopedia Británica* conoce un individuo determinado; pero esto no nos daría ninguna medida de los hechos conocidos por el individuo, pero que no están contenidos en la enciclopedia. El método sólo es útil para establecer un límite inferior.

Un experimento relacionado con éste es el juego de las veinte preguntas, en el que uno de los jugadores identifica un objeto escogido por otro planteando una serie de veinte preguntas por sí o por no. Dado que cada respuesta proporciona no más que un bit de información, y dado que los jugadores habilidosos necesitan ge-neralmente plantear casi todas las veinte preguntas para

identificar correctamente el objeto escogido, podemos estimar que el número de elecciones permitidas conocidas en común por dos jugadores es del orden de 2^{20}, o alrededor de un millón. Por supuesto, esta medida es inadecuada porque las preguntas no son perfectas y porque la elección de los objetos no es al azar. Es posible que se pueda desarrollar una versión refinada del juego, y que se la pueda utilizar para proporcionarnos otro límite inferior.

Una tercera estrategia para medir los requerimientos de información en la porción simbólica del conocimiento del cerebro humano es estimar la tasa promedio de adquisición de información y calcular la cantidad que se acumularía con el tiempo. Por ejemplo, los experimentos sobre memorización de secuencias de sílabas al azar indican que la tasa máxima de memorización de este tipo de conocimiento es de alrededor de un «chunk» por segundo. Se puede suponer que un «chunk», en este contexto, contiene menos de 100 bits de información, de modo que los resultados sugieren que la tasa máxima a la cual un humano es capaz de depositar información en la memoria a largo plazo es significativamente menor a los 100 bits por segundo.[2] Si esto es verdad, el aprendizaje de un humano de veinte años de edad a una tasa máxima para seis horas por día (y no olvidando nunca) sería de menos de 50 mil millones de bits de información. Encuentro este número sorprendentemente pequeño.

Una dificultad con esta estimación de la tasa de adquisición es que el experimento mide sólo información que proviene de un solo canal sensorial bajo un conjunto particular de circunstancias. El sistema visual envía más de un millón de veces esta tasa de información al nervio óptico, y es concebible que toda esa información se deposite en la memoria. Si llega a ser cierto que las imágenes se almacenan directamente, será necesario incrementar significativamente el límite de 100 bits por segundo, pero no hay evidencia actual de que éste sea el caso. En experimentos que miden la habilidad de individuos excepcionales para almacenar imágenes eidéticas (es decir, extraordinariamente exactas y vívidas) de estereogramas de puntos al azar, se da a los sujetos alrededor de cinco minutos para memorizar una imagen formada en una matriz cuadrada de 100 x 100 puntos. Memorizar sólo unos pocos cientos de bits es probablemente suficiente para pasar la prueba.

Me temo que no hay ninguna evidencia que sugiera que se puede depositar en la memoria a largo plazo más que unos pocos

bits por segundo de cualquier tipo de información. Aun si aceptamos sin crítica reportes de hazañas de memoria extraordinarias (tales como la del showman de Luria en *Mind of the Mnemonist*[3]), la tasa promedio de depósito en memoria nunca parece exceder unos pocos bits por segundo; los experimentos deberían refinar esta estimación. Aun si supiéramos exactamente la tasa máxima de memorización, la tasa promediada a lo largo de una vida probablemente sería mucho menor; pero conociendo la tasa máxima estableceríamos un límite superior a los requerimientos de almacenamiento.

Los datos sucintos mencionados arriba sugieren que una máquina inteligente requeriría 10^9 bits de almacenamiento, más o menos dos órdenes de magnitud. Esto supone que la información se codifica de tal forma que requiere una cantidad mínima de almacenamiento; para el propósito de procesar información, ésta probablemente no sería la representación más práctica. Como un constructor potencial de máquinas pensantes, encuentro que este número es alentadoramente pequeño, dado que se encuentra dentro del rango de las computadoras electrónicas actuales. Como seres humanos con un ego, encuentro que esto es descorazonador: no me gusta pensar que toda una vida de recuerdos se puede colocar sobre un rollo de cinta magnética. Es de esperar que la evidencia experimental aclarará esto de un modo o de otro.

Hay unas pocas sutilezas en el problema de las exigencias de almacenamiento que involucran definir la cantidad de información de una manera que sea independiente de su representación. La teoría de la información proporciona una forma precisa de medir la información en términos de bits, pero esto requiere una medida de las probabilidades sobre el conjunto de los estados posibles. Esto es, requiere asignar una probabilidad a priori a cada conjunto posible del conocimiento, y éste es el papel de la inteligencia heredada. La inteligencia heredada proporciona un marco de referencia en el cual se puede interpretar el conocimiento de la inteligencia adquirida. La inteligencia heredada define lo que es conocible; la inteligencia adquirida determina qué es lo que se conoce de lo conocible.

Otra dificultad potencial es cómo contar el almacenamiento de la información que puede ser deducida de otros datos. En el sentido estricto de la teoría de la información, los datos que pueden inferirse de otros no agregan en absoluto información. Una medición adecuada no tomaría en cuenta la posibilidad de

que el conocimiento sea inconsistente, y de que sólo se realizan en realidad inferencias limitadas. Estos son tipos de cuestiones que se están estudiando en el bando simbólico del campo de la inteligencia artificial.

Una cuestión que no necesita estar resuelta para poder medir la capacidad de almacenamiento es la de la representación localizada versus la representación distribuida: es decir, si cada pieza de información se almacena en un lugar específico o si se desparrama «holográficamente» sobre una gran área. Saber qué tipos de representación se utilizan en qué partes del cerebro humano es de considerable interés científico, pero no posee un impacto profundo sobre la cantidad de almacenamiento en el sistema o sobre nuestra capacidad para medirla. Los comentaristas no técnicos tienen tendencia a atribuir cualidades casi místicas a los mecanismos de almacenamiento distribuido, como los que se usan para crear hologramas y redes neuronales, pero las limitaciones de las capacidades de estos mecanismos de almacenamiento se comprenden muy bien.

Cuando una placa holográfica se corta en dos, cada una contiene una imagen ligeramente degradada de la imagen completa. Las representaciones distribuidas con propiedades similares a las de los hologramas se utilizan a menudo en computadoras digitales convencionales, y son invisibles a la mayoría de los usuarios, excepto en la capacidad del sistema para tolerar errores. Los sistemas de corrección de memoria que se utilizan en casi todas las computadoras son un buen ejemplo. El sistema se compone de muchos chips de memoria físicamente separados, pero se puede quitar cualquier chip sin perder ningún dato. Esto se debe a que los datos no se almacenan en un solo lugar, sino en una representación distribuida, no local, a través de todas las unidades. A despecho de esta representación «holográfica», la capacidad de almacenamiento de información del sistema no es mayor de lo que sería con una representación convencional, en la que cada pieza de dato se almacenara en un solo chip. De hecho, es ligeramente menor. Esto es típico de las representaciones distribuidas.

La capacidad de almacenamiento ofrece una medida de los requerimientos de una inteligencia emergente similar a la humana. Otra medida es la tasa de computación. Aquí no hay ninguna métrica consensuada, y es particularmente difícil definir una unidad de medida que sea completamente independiente de la

representación. La medida sugerida más adelante es simple e importante, si es que no suficiente.

Dada una representación eficientemente almacenada del conocimiento humano, ¿qué tasa de acceso a ese almacenamiento (en bits por segundo) se requiere para alcanzar una performance semejante a la humana? Aquí, *representación eficientemente almacenada* significa cualquier representación que requiera sólo una constante multiplicativa de almacenamiento sobre el número de bits de información. Esta es una restricción matemática que elimina, por ejemplo, cualquier representación que almacene una respuesta precomputada a cada pregunta. Esa restricción limita el rango de representaciones posibles, pero permite la mayoría de las representaciones que consideraríamos razonables. En particular, permite tanto representaciones distribuidas como locales.

La cuestión del ancho de banda de memoria requerida para una performance semejante a la humana es accesible experimentalmente mediante estrategias similares a las referidas a propósito de la cuestión de la capacidad de almacenamiento. Si el tiempo requerido para una operación primitiva de la memoria humana se limita al tiempo de disparo de una neurona, la división entre este «tiempo de ciclo» y el número total de bits indica la fracción de memoria a la que se accede simultáneamente. Esto proporciona un indicio respecto de si el cerebro es un dispositivo paralelo o secuencial. En un dispositivo serial, los ítems de datos se operan secuencialmente, uno por vez. En un dispositivo paralelo, todos los datos se operan concurrentemente. El cerebro exhibe tanto conductas seriales como paralelas, pero hay una cuestión respecto a cuál de los modelos describe mejor la forma en que se razona y se accede al conocimiento. Las opiniones fundadas difieren enormemente a este respecto, pero el núcleo de la evidencia cuantitativa favorece a la computación serial. Los tiempos de recuperación de la memoria para elementos en una lista, por ejemplo, dependen de la posición y de la ubicación de los elementos en la lista. Excepto para el procesamiento sensorial, los programas más exitosos de inteligencia artificial se han basado en modelos seriales de computación, aunque esto puede ser una distorsión causada por la disponibilidad común de máquinas seriales.

Mi propia conjetura es que los experimentos de tiempo de reacción son engañosos y que la performance a nivel humano requerirá acceso a vastas fracciones de conocimientos varias

veces por segundo. Dada una representación de inteligencia adquirida con una eficiencia de representación realista del 10 por ciento, los 10^9 bits de memoria antes mencionados requerirían un ancho de banda de unos 10^{11} bits por segundo. Este ancho de banda parece fisiológicamente plausible, dado que corresponde a cerca de un bit por segundo por neurona en la corteza cerebral.

En comparación, el ancho de banda de memoria de una computadora secuencial convencional está en el rango de los 10^6 a 10^8 bits por segundo. Esto es menos que el 0,1 por ciento del requerimiento imaginado. Para las computadoras en paralelo el ancho de banda es considerablemente mayor. Por ejemplo, una Máquina de Conexión de 65.536 procesadores puede acceder a su memoria a aproximadamente 10^{11} bits por segundo.[4] No es completa coincidencia que esto encaje bien con lo estimado arriba.

Otra cuestión importante es: ¿Qué funciones sensorio-motoras se necesitan para sustentar la inteligencia simbólica? Un simio es una máquina sensorio-motora compleja, y es posible que gran parte de esta complejidad sea necesaria para sustentar la inteligencia. Grandes porciones del cerebro parecen dedicadas al procesamiento visual, auditivo y sensorial, y se desconoce cuánta de toda esta maquinaria se necesita para pensar. Una persona ciega, sorda o totalmente paralizada puede sin duda ser inteligente, pero esto no prueba que la porción del cerebro dedicada a estas funciones sea innecesaria para el pensamiento. Pudiera ser, por ejemplo, que un ciego sacara ventaja del aparato de procesamiento visual del cerebro para el razonamiento espacial.

A medida que comencemos a comprender mejor la arquitectura funcional del cerebro, sería posible identificar ciertas funciones como innecesarias para el pensamiento estudiando pacientes cuyas habilidades cognitivas no estén afectadas por daños cerebrales localmente confinados. Por ejemplo, se sabe que la fusión binocular estereoscópica tiene lugar en una región específica de la corteza cerca de la parte trasera de la cabeza. Los pacientes con daños en esta área de la corteza tienen desventajas visuales, pero no muestran una caída evidente en su habilidad para pensar. Esto sugiere que la fusión estereoscópica no es necesaria para el pensamiento. Este es un ejemplo simple, y la conclusión no es sorprendente, pero sería posible que esos experimentos establecieran que muchas funciones sensorio-motoras son innecesarias. Nos podemos imaginar cercenando metafóricamente el cerebro

hasta reducirlo a su núcleo esencial. Por supuesto, esto no es tan simple. El daño cerebral raramente incapacita una sola área del cerebro total y exclusivamente. También puede ser difícil eliminar una función a la vez, porque una función mental bien puede compensar la falta de otra.

Puede ser más productivo suponer que todo el aparato sensorio-motor es innecesario hasta que se pruebe su utilidad para el pensamiento, pero esto es contrario al punto de vista usual. Nuestra actual comprensión del desarrollo filogenético del sistema nervioso sugiere un punto de vista en el cual la inteligencia es un refinamiento elaborado de la conexión entre entradas y salidas. Esto se halla reforzado por la conveniencia experimental de estudiar sistemas nerviosos simples, o de estudiar sistemas nerviosos complejos concentrándonos en las porciones más directamente relacionadas con las entradas y salidas. Por necesidad, casi todo lo que sabemos de la función del sistema nervioso proviene de experimentos sobre esas porciones que están estrechamente relacionadas con los insumos sensoriales y las respuestas motoras. No sería sorprendente averiguar que hemos sobrestimado la importancia de esas funciones en el pensamiento inteligente.

Las funciones sensorio-motoras son claramente importantes para la aplicación de la inteligencia y para su evolución, pero estas cuestiones están separadas del problema de si las funciones sensorio-motoras son necesarias para que el pensamiento exista. La inteligencia no sería de mucha utilidad sin un sistema elaborado de aparatos sensoriales para medir el entorno y un sistema elaborado de aparatos motores para cambiarlo, ni habría sido probable que evolucionara. Pero probablemente se necesita mucho más aparato para ejercitar y desarrollar la inteligencia que para sustentarla. Se puede creer en la necesidad del pulgar oponible para el desarrollo de la inteligencia sin dudar de una capacidad humana para el pensamiento sin pulgar. Es harto posible que incluso las magras capacidades sensorio-motoras que hoy sabemos crear artificialmente sean suficientes para la operación fundamental de la inteligencia emergente.

Aunque las cuestiones de capacidad y amplitud son necesarias para definir la magnitud de la tarea de construir una inteligencia emergente, la cuestión clave es la de la comprensión. Aunque es posible que seamos capaces de recrear el sustrato emergente de

la inteligencia sin comprender totalmente los detalles de la forma en que trabaja, resulta probable que necesitemos al menos comprender algunos de sus principios. Hay por lo menos tres vías para alcanzar esa comprensión. Una consiste en estudiar las propiedades de sistemas emergentes específicos: construir una teoría de sus capacidades y limitaciones. Esta clase de estudio experimental se está llevando a cabo sobre diversas clases de sistemas hechos por el hombre, incluyendo redes neuronales, cristales de spin, autómatas celulares, sistemas evolutivos y autómatas adaptativos. Otra vía posible hacia la comprensión es el estudio de los sistemas biológicos, que son nuestros únicos ejemplos reales de inteligencia y nuestros únicos ejemplos de un sistema emergente que ha producido inteligencia. Las disciplinas que hasta hoy han proporcionado la información más útil de este tipo han sido la neurobiología, la psicología cognitiva y la biología evolutiva. Una tercera vía sería la comprensión teorética de los requerimientos de la inteligencia o del fenómeno de la emergencia. Ejemplos relevantes son las teorías de la lógica y la computabilidad, la lingüística y la teoría de los sistemas dinámicos. Cualquiera que contemple los sistemas emergentes como formas de defender el pensamiento humano del escrutinio de la ciencia probablemente quede defraudado.

No podemos concluir, sin embargo, que sea necesaria una comprensión reduccionista para la creación de inteligencia. Incluso una modesta comprensión puede implicar un largo trecho para la construcción de un sistema emergente. Un buen ejemplo de esto es la forma en que se han usado autómatas celulares para simular la conducta emergente de los fluidos. Los remolinos que se forman cuando un fluido pasa a través de un obstáculo no están bien comprendidos analíticamente, pero son de gran importancia práctica para el diseño de barcos y aviones. Las ecuaciones que describen el flujo de un fluido se han conocido durante casi un siglo, pero excepto en pocos casos no se han podido resolver. En la práctica el fluido se analiza generalmente mediante la simulación. El método de simulación más común es la solución numérica de ecuaciones continuas.

En una computadora masivamente paralela es posible simular fluidos con una comprensión del sistema aún menor, simulando billones de partículas en choque que reproducen fenómenos emergentes, tales como los vórtices. Calcular las interacciones

moleculares detalladas de tantas partículas sería extremadamente difícil, pero unos pocos aspectos simples del sistema, tales como la conservación de energía y el número de partículas, son suficientes para reproducir el comportamiento a gran escala. Un sistema de partículas simplificadas que obedecen estas dos leyes pero que no son nada realistas en otros aspectos, puede reproducir los mismos fenómenos emergentes que la realidad. Por ejemplo, es posible usar partículas de masa unitaria que se mueven sólo a velocidades unitarias a lo largo de una matriz hexagonal, chocando de acuerdo con las reglas de las bolas de billar.[5] Los experimentos muestran que este modelo produce flujo laminar, corrientes en vórtice e incluso turbulencias que son indistinguibles de la conducta de los fluidos reales. Aunque las reglas detalladas de la interacción son muy diferentes de las de las interacciones de las moléculas reales, los fenómenos emergentes son los mismos. Se pueden crear fenómenos emergentes sin comprender los detalles de las fuerzas entre las moléculas o las ecuaciones que describen el flujo de los fluidos.

La recreación de patrones intrincados de mareas y flujos en un fluido demuestra que es posible producir un fenómeno sin comprenderlo plenamente. Pero el modelo fue construido por físicos que sabían una enormidad sobre fluidos. Ese conocimiento ayudó a determinar cuáles rasgos del sistema físico era importantes para implementar y cuáles no.

La física es una ciencia excepcionalmente exacta. Quizás un mejor ejemplo de un sistema emergente que podemos simular con apenas una comprensión limitada es la biología evolucionaria. Comprendemos, en un sentido muy débil, la forma en que criaturas con patrones mendelianos de herencia y diferentes propensiones para la supervivencia pueden desarrollarse en pos de una mejor adecuación con sus entornos. En ciertas situaciones simples incluso podemos escribir ecuaciones que describen la velocidad con que tendrá lugar esa evolución.[6] Pero en nuestra comprensión de los procesos evolutivos hay muchas lagunas. Podemos explicar por qué los animales que vuelan tienen huesos livianos en términos de la selección natural, pero no podemos explicar por qué algunos animales han desarrollado vuelo y otros no. Tenemos alguna comprensión cuantitativa de las fuerzas que causan el cambio evolutivo, pero (excepto en ciertos casos simples) no podemos explicar la tasa ni aun la dirección de ese cambio.

A despecho de estas limitaciones, nuestra comprensión es suficiente para escribir programas de evolución simulada que muestran interesantes conductas emergentes. Por ejemplo, recientemente he utilizado una simulación evolutiva para desarrollar programas que ordenan números. En este sistema, el material genético de cada individuo simulado se interpreta como un programa que especifica un patrón de comparaciones e intercambios. La probabilidad de supervivencia de un individuo en el sistema depende de la eficiencia y adecuación de su programa de ordenamiento de números. Los individuos que sobreviven producen descendencia mediante combinación sexual de su material genético con ocasionales mutaciones al azar. Después de decenas de miles de generaciones, una población de cientos de miles de estos individuos desarrollará programas de ordenamiento muy eficientes. Aunque escribí la simulación que produce estos programas de ordenamiento, no comprendo en detalle cómo fueron producidos ni cómo trabajan. Si la simulación no hubiera producido programas operativos, tendría muy poca idea de cómo arreglarla.

Los ejemplos del flujo de los fluidos y de la evolución simulada sugieren que es posible hacer mucho con una pequeña cantidad de comprensión. Las conductas emergentes exhibidas por estos sistemas son una consecuencia de las simples reglas subyacentes definidas por el programa. Aunque los sistemas tienen éxito para producir los resultados deseados, están más allá de nuestra habilidad para analizar y predecir. Se puede imaginar que si un proceso similar produjera un sistema de inteligencia emergente, tendríamos una falta de comprensión similar sobre su forma de trabajo.

Mi propia conjetura es que tal sistema emergente no será en sí mismo un sistema inteligente, sino más bien el sustrato metabólico a partir del cual la inteligencia podrá crecer. En término de los simios y las canciones, la porción emergente del sistema jugará el papel del simio, o al menos de la parte del simio que alberga las canciones. Esta mente artificial necesitará ser inoculada con conocimiento humano. Imagino que este proceso no es muy diferente de la educación de un niño. Este sería un procedimiento riesgoso e incierto porque, igual que un niño, esta mente emergente será susceptible presumiblemente tanto a las ideas malas como a las buenas. El resultado no será tanto una inteligencia artificial, sino una inteligencia humana contenida en una mente artificial.

Por supuesto, comprendo que esto es solamente un sueño y admitiré que me he dejado llevar más por la esperanza que por la probabilidad del éxito. Pero si esta mente artificial puede sustentarse y crecer por sus propios medios, entonces el pensamiento humano vivirá por primera vez libre de huesos y carnes, dando a este hijo de la mente la inmortalidad terrenal que nos ha sido negada.

Notas

[1] Freeman Dyson, *The Origins of Life* (Cambridge: Cambridge University Press, 1985).

[2] Allen Newell, *Human Problem Solving* (Englewood Cliffs, N. J. : Prentice Hall, 1972).

[3] A. R. Luria, *Mind of the Mnemonist* (Nueva York: Basic Books, 1968).

[4] Daniel W. Hillis, *The Connection Machine* (Cambridge: MIT Press, 1985).

[5] Stephen Wolfram, *Theory of Applications of Cellular Automata* (World Scientific, 1986).

[6] J. B. S. Haldane, *The Causes of Evolution* (Harper & Brothers, 1932).

9
Perspectivas de la construcción de máquinas verdaderamente inteligentes

David L. Waltz

¿Puede alcanzarse la inteligencia artificial? Si es así, ¿en cuánto tiempo? ¿Mediante qué métodos? ¿Qué ideas de la IA contemporánea resultarán importantes a largo plazo para una ciencia de la cognición? Creo que la IA podrá alcanzarse, quizás en el curso de nuestras vidas, pero debemos sortear importantes obstáculos científicos y de ingeniería para lograrlo. Los métodos y la perspectiva de la IA han resultado dramáticamente sesgados por la existencia de la computadora digital común, llamada a veces la máquina de von Neumann; y en última instancia la IA tendrá que basarse en ideas y en hardware harto distintos de lo que actualmente es central en ella. La memoria, por ejemplo, es mucho más importante de lo que sugiere su papel actual en la IA, y la búsqueda de datos tiene mucha menos importancia que la que le hemos adjudicado. Además, dado que las computadoras carecen de cuerpo y de experiencias de vida similares a las de los humanos, los sistemas inteligentes serán tal vez inherentemente distintos de los humanos; especulo brevemente a qué se parecerán esos sistemas.

Obstáculos para construir sistemas inteligentes

Si hemos de construir máquinas que sean tan inteligentes como la gente, tenemos que resolver tres problemas: debemos establecer una ciencia de la cognición, debemos diseñar el softwa-

---

David L. Waltz. Profesor de computación científica en la Universidad Brandeis, científico principal y director de sistemas avanzados de información en la Thinking Machines Corporation.

re, los sensores y los efectores para un sistema completo, y debemos diseñar el hardware adecuado.

Estableciendo una ciencia de la cognición

No tenemos una ciencia de la cognición adecuada. Sólo tenemos fragmentos de la concepción, y algunos de ellos son por cierto incorrectos. Sabemos muy poco sobre la forma en que deberá organizarse la máquina para resolver los problemas de la inteligencia. Virtualmente todos los aspectos de la inteligencia (incluyendo la percepción, la memoria, el razonamiento, la intención, la generación de acciones y la atención) son todavía misteriosos. Sin embargo, aun si hubiéramos comprendido la forma de estructurar un sistema inteligente, no seríamos capaces de completarlo porque carecemos de una ciencia del conocimiento apropiada. Para algunos aspectos del conocimiento, cualquier dispositivo computacional será un fundamento fuerte en comparación con una persona. Las enciclopedias, los diccionarios y los textos legibles por máquinas, con el tiempo, permitirán que las máquinas absorban conocimiento libresco con rapidez. Para que esa comprensión sea profunda, sin embargo, un sistema necesita bases perceptuales y una comprensión del mundo físico y social. Para los humanos, gran parte de ese conocimiento o bien es innato o está organizado y reunido por estructuras innatas que automáticamente nos hacen prestar atención a ciertos rasgos de nuestra experiencia que entonces consideramos importantes. Será extremadamente difícil construir dentro de un sistema las clases de conocimiento a priori o de principios estructurales que poseen los humanos.

Diseñando el software

Cualquier sistema verdaderamente inteligente deberá ser grande y complejo. Como argumenta Frederick Brooks, escribiendo sobre su experiencia en la construcción del gran sistema operativo OS360 en IBM, no es posible acelerar un proyecto de software simplemente poniendo en él más y más gente.[1] El tamaño óptimo del equipo para construir el software es de alrededor de cinco personas. Por esta razón, y debido a la delicada envergadura de un proyecto de esta clase (que empequeñece a cualquier otro que se

haya intentado hasta ahora en programación), la codificación a mano será por cierto demasiado lenta y poco confiable para coronar toda la tarea. En consecuencia, un sistema verdaderamente inteligente deberá ser capaz de aprender de la experiencia gran parte de su estructura.

¿Qué estructuras deberán construirse dentro del sistema para permitirle aprender? Esta es una cuestión central para la IA actual, y la respuesta depende de cuestiones de representación del conocimiento: ¿Cómo se deberá representar el conocimiento? ¿Con qué componentes (si es que con alguno) están construidas las estructuras del conocimiento?

Creando el hardware

Debemos poder construir hardware adecuado a las necesidades de la representación del conocimiento y el aprendizaje en IA y que se compare en poder con el cerebro humano. Nadie debe sorprenderse de que las diminutas máquinas que la IA ha utilizado hasta hoy no hayan exhibido inteligencia artificial. Incluso las computadoras actuales más poderosas son no más que un cuatrillonésimo de poderosas que la mente humana. Además, las máquinas actuales son probablemente al menos igual de deficientes en capacidad de memoria: las computadoras actuales más grandes no tienen más que un cuatrillonésimo de la capacidad de memoria del cerebro humano. Aun dadas estas extremas discrepancias, el hardware tal vez demuestre ser la parte más fácil de la tarea que debe desarrollar la IA.

Comienzo ahora con una discusión de la IA tradicional y de sus limitaciones teóricas para establecer el escenario para una discusión de los giros (o quiebras) paradigmáticos principales que actualmente se manifiestan en la IA. Como un defensor de la necesidad de nuevos paradigmas, confieso aquí mi parcialidad. No veo modo de que los métodos tradicionales de la IA se puedan extender para alcanzar una inteligencia similar a la humana. Suponiendo que nuevos paradigmas reemplazarán o se combinarán con los tradicionales, hago algunas proyecciones sobre cuánto falta para que se puedan construir sistemas inteligentes y a qué se parecerán.

Límites de la IA tradicional

Dos giros paradigmáticos revolucionarios se están manifestando en la inteligencia artificial. La fuerza principal que está detrás de los giros es la creciente sospecha entre los investigadores de que los modelos actuales de IA no parece que puedan ampliarse hasta el punto en que manifiesten una inteligencia parecida a la humana. Los giros apuntan hacia computadoras masivamente paralelas y hacia programas masivamente paralelos que se enseñan, más que se programan. Los sistemas resultantes de hardware y software se parecen en muchos aspectos más a los cerebros que a las máquinas seriales de von Neumann y a los programas a los que nos hemos acostumbrado.

Durante treinta años, virtualmente todos los paradigmas de IA se basaron en variantes de lo que Herbert Simon y Allen Newell han presentado como un «sistema físico de símbolos» y la hipótesis de la «búsqueda heurística».[2] (Véase también el artículo de Hubert y Stuart Dreyfus en esta edición de *Dædalus*.)

De acuerdo con la hipótesis del sistema físico de símbolos, los símbolos (entidades parecidas a las palabras o a los números, los nombres de los objetos y sucesos) son los objetos primitivos de la mente; mediante algún proceso desconocido, el cerebro imita una «máquina de inferencia lógica», cuyo rasgo más importante es que es capaz de manipular símbolos (es decir, recordar, interpretar, modificar, combinar y expandir a partir de ellos), y los modelos de computadora que capturan símbolos capturan por lo tanto las operaciones esenciales de la mente. En esta argumentación no importa si los materiales con que está construida esta máquina de inferencia son transistores o neuronas. La única cosa importante es que sean capaces de un conjunto universal de operaciones lógicas.[3] La hipótesis del sistema físico de símbolos a su vez descansa en una fundamentación de resultados matemáticos sobre computabilidad, que puede utilizarse para demostrar que si una máquina es equivalente a una máquina de Turing (una forma simple de modelo computacional desarrollado por el pionero matemático británico Alan Turing) entonces es «universal»; es decir, la máquina puede computar cualquier cosa que pueda ser computada. Se puede demostrar que todas las computadoras digitales ordinarias son universales en el sentido de Turing. (Este es quizás un aspecto crítico en el cual todas las computadoras son

incapaces de equipararse a una máquina de Turing: la máquina de Turing incluye una cinta infinita, de la que lee sus programas y en la que escribe sus resultados. Todas las computadoras —y presumiblemente todos los humanos— poseen una memoria finita.)

En el modelo de la búsqueda heurística, los problemas de la cognición son instancias de la exploración de un espacio de posibilidades para una solución. El espacio de búsqueda para los problemas de búsqueda heurística se puede visualizar como un árbol que se ramifica: comenzando en la raíz del árbol, cada alternativa considerada y cada decisión que se toma corresponden a un punto de ramificación del árbol. Las heurísticas, o las reglas empíricas, permiten que la búsqueda se concentre primero en las ramas en las que es probable que haya una solución, impidiendo así una búsqueda combinatoriamente explosiva a través de todo el espacio de soluciones. (Los problemas combinatoriamente explosivos son problemas en los que los costos computacionales para resolver cada problema ligeramente más difícil crecen tan rápidamente que ninguna computadora podría ser capaz de resolverlo; es decir, incluso una computadora con tantos componentes como electrones haya en el universo y un tiempo de ejecución tan breve como el más breve suceso físico susceptible de medirse requerirá tiempos mayores a la edad del universo para considerar todas las soluciones posibles del problema.) Los problemas de búsqueda heurística son fáciles de implementar en computadoras digitales ordinarias. La búsqueda heurística se ha utilizado para una amplia variedad de aplicaciones, incluyendo toma de decisiones, juegos, planeamiento de robots y resolución de problemas, procesamiento del lenguaje natural y la clasificación de objetos perceptuales. La búsqueda heurística ha disfrutado de particular prominencia, porque está en el corazón de los «sistemas expertos», con mucho el mayor éxito comercial de la IA.

En retrospectiva, es llamativa la seriedad con que se ha tomado la búsqueda heurística como modelo cognitivo. Cuando yo era estudiante graduado a fines de la década de 1960, la concepción estándar de la IA era que, para cualquier sistema inteligente, la naturaleza del problema limitaba la naturaleza de cualquier solución eficiente, y que cualquier sistema, humano o computadora, dado un problema a resolver, tendía a desarrollar una estructura interna similar, o por lo menos análoga, para tratar con él. De esta forma, se argumentaba, estudiar soluciones eficientes

de problemas en las computadoras era una buena forma de estudiar la cognición.[4] Virtualmente todos en la IA de esa época aceptaban la centralidad y la inmutabilidad de la maquinaria de búsqueda heurística sin cuestionarla y suponiendo que el aprendizaje se llevaría a cabo desarrollando, adaptando o agregando a las heurísticas y a las estructuras de conocimiento del espacio de búsqueda. (Las excepciones eran los investigadores de «redes neuronales» y «perceptrones», quienes habían explorado activamente modelos más parecidos al cerebro desde comienzos de la década de 1950. Más sobre esto más adelante.)

Ahora se reconoce habitualmente que la naturaleza de las computadoras y de los modelos computacionales de que disponemos inevitablemente limitan los algoritmos de resolución de problemas que podemos considerar. (John Backus introdujo esta idea a la amplia comunidad científica en su conferencia del Premio Turing de 1977[5].) Como se explica más adelante, se tornó evidente que los métodos tradicionales de la IA no escalaban bien y que por lo tanto se necesitaban nuevos paradigmas. A despecho de este cambio de actitud, ha habido pocos reemplazos prospectivos dentro de la IA para la búsqueda heurística (o para las computadoras seriales de un solo procesador) hasta hace muy poco.

Las razones por las que la IA se ha concentrado casi exclusivamente en torno al sistema físico de símbolos y a la búsqueda heurística se hallan profundamente enraizadas en su historia y reflejan en parte la miope concentración alrededor de las computadoras digitales seriales que ha caracterizado a toda la computación científica. La concentración en torno a la búsqueda heurística también refleja la influencia de la investigación psicológica de la década de 1950. La IA comenzó en una época en que los psicólogos se hallaban enamorados del análisis de protocolos, una forma de examinar la conducta humana haciendo que los sujetos dieran cuenta de su experiencia mental mientras resolvían problemas.[6] Dicha investigación psicológica se interpretaba como evidencia de que el principal mecanismo humano para la resolución de problemas es el ensayo y error. La IA adoptó este modelo como su paradigma de búsqueda heurística. En este paradigma los problemas se resuelven aplicando secuencialmente «operadores» (pasos elementales en la solución de un problema) y permitiendo la «búsqueda hacia atrás» (*backtracking*), una forma de ensayo y error en la que un programa vuelve a un punto anterior

de decisión e intenta nuevas ramas si las primeras que ha explorado demuestran no ser fructíferas.

Es difícil ver cómo cualquier extensión de los sistemas basados en la búsqueda heurística podría alguna vez demostrar sentido común. En la mayoría de los sistemas de IA, los enunciados de un problema provienen de los usuarios; los sistemas no han decidido sobre qué problemas trabajar. Estos tienen relativamente pocas acciones u operadores disponibles, para que los espacios de búsqueda sean tratables. La performance en tiempo real por lo general no ha sido necesaria. Esta forma de operar claramente no funcionará. Con el tiempo, la IA debe enfrentarse al problema del escalamiento: dado el inmenso rango de situaciones en que se puede encontrar un sistema verdaderamente inteligente y el vasto número de acciones posibles al alcance, ¿cómo puede el sistema arreglárselas para buscar los objetivos y las acciones apropiados?

Además, como lo ha señalado John McCarthy, los sistemas basados en reglas pueden hallarse inherentemente limitados por el «problema de la cualificación»: dada una cierta regla general, siempre se puede alterar la situación del mundo de tal manera que la regla ya no sea apropiada.[7] Por ejemplo, supongamos que hemos ofrecido esta regla:

ave(x) $\longrightarrow$ *vuela(x)* (si *x* es un ave, entonces *x* puede volar).

Todo el mundo sabe que la regla debe enmendarse para cubrir aves como los pingüinos y los avestruces, de modo que se transforma en:

no sinvuelo(x) y ave(x) $\longrightarrow$ *vuela(x)*

donde «sinvuelo(*x*)» es verdad para las aves apropiadas.

Sin embargo, también sabemos que un ave no puede volar si está muerta, o si se le han cortado las alas, o si sus pies se han impregnado con cemento, o si ha sido condicionada dándole choques eléctricos cada vez que trata de volar.[8] No parece haber forma de especificar por completo las reglas para esos casos. También hay serias dificultades para formular reglas para decidir cuáles son los hechos del mundo que se han de retirar y cuáles deben mantenerse después que han ocurrido sucesos o acciones

particulares. Esto se conoce como el «problema del marco» (*frame*). Para tratar estos problemas se ha propuesto la «lógica no monotónica», que trata todas las reglas o proposiciones nuevas como hipótesis retirables.[9] Sin embargo, algunos investigadores en esta área[10] son pesimistas sobre su potencial, y yo lo soy.

Al objetar las estrategias tradicionales de IA, no estoy discutiendo las nociones de computación universal o los resultados de la máquina de Turing, que están establecidos matemáticamente más allá de toda duda. Más bien, discuto la metáfora de la búsqueda heurística, la relación entre sistemas físicos de símbolos y cognición humana, y la naturaleza y «granularidad» de las unidades de pensamiento. La hipótesis de los sistemas físicos de símbolos, compartida desde hace tiempo por los investigadores en IA, afirma que un vocabulario próximo al lenguaje natural (el inglés, por ejemplo, previamente suplementado por categorías y conceptos que antes no tenían nombre) sería suficiente para expresar todos los conceptos que necesitaran expresarse. Mi creencia es que los términos parecidos a los del lenguaje natural son, para algunos conceptos, burdos y ambiguos sin esperanza, y que deben hacerse muchas distinciones más finas, «subsimbólicas», especialmente para codificar los insumos sensoriales. Al mismo tiempo, algunas unidades mentales (por ejemplo, situaciones o sucesos totales, a menudo recordados como imágenes mentales) parecen ser importantes portadoras de significado que no pueden reducirse a estructuras tratables de palabras o de entidades similares a las palabras. Todavía peor, creo que las palabras no son en ningún caso portadoras de significados plenos, sino más bien como términos de indicio o pistas que un hablante utiliza para inducir a un oyente a extractar recuerdos y conocimientos compartidos. El grado de detalle y el número de unidades que se necesitan para expresar el conocimiento y la intención del hablante y la comprensión del oyente son vastamente mayores que el número de palabras que se usa para comunicarse. En este sentido, el lenguaje puede ser como el juego de las charadas: el hablante transmite relativamente poco, y el oyente genera comprensión mediante la síntesis de los elementos de memoria evocados por los indicios del hablante. En forma parecida, creo que las palabras que parecen ampliamente características de las corrientes de conciencia humanas no constituyen en sí mismas pensamiento; más bien, representan una proyección de nuestros

pensamientos sobre nuestras facultades de producción de habla. De este modo, por ejemplo, podemos sentirnos felices o confundidos sin haber jamás formado esas palabras o podemos resolver un problema imaginando un diagrama sin palabras o con muy pocas palabras especificar un diagrama.

¿Cuál es la alternativa?

En otra parte, Craig Stanfill y yo argumentamos extensamente que los humanos bien pueden resolver problemas mediante un proceso mucho más parecido a *mirar* que a *buscar*, y que los elementos mirados pueden ser más parecidos a representaciones de episodios y objetos específicos o estereotipados que a las reglas y los hechos.[11]

En la Máquina de Conexión, construida por la Thinking Machines Corporation,[12] hemos implementado varios tipos de sistemas de «memoria asociativa» que razonan sobre la base de la experiencia previa.[13] Por ejemplo, un sistema experimental resuelve problemas de diagnóstico médico con «razonamiento basado en los recuerdos»: dado un conjunto de síntomas y características del paciente, el sistema encuentra los pacientes previos más parecidos y lanza la hipótesis de que el mismo diagnóstico se ha de aplicar al paciente nuevo. Los modelos «conexionistas» o redes neuronales, que describiré más adelante, resuelven problemas similares, aunque en forma muy distinta. Pese a que todavía se requiere mucha investigación para que estos sistemas sean candidatos serios para ser sistemas verdaderamente inteligentes, creo que estas arquitecturas demostrarán ser mucho más fáciles de construir y de extender que los modelos de búsqueda heurística. Estos nuevos modelos pueden aprender y razonar recordando y generalizando ejemplos específicos; los modelos de búsqueda heurística, por el contrario, dependen de reglas. Se ha demostrado que es difícil obtener reglas de los expertos, quienes a menudo ni siquiera advierten que están usando reglas. No sabemos cómo verificar la completud y la autoconsistencia de conjuntos de reglas. Además, un conjunto finito de reglas no puede capturar todas las conclusiones posibles que se pueden extraer de un conjunto de ejemplos, más de lo que un conjunto de frases descriptivas puede describir completamente una imagen.

Es importante notar, sin embargo, que algunas clases de conocimiento en los sistemas basados en reglas son difíciles de codificar en nuestros modelos basados en recuerdos. Por ejemplo, como se lo formula actualmente, nuestro sistema no utiliza historias de pacientes y es incapaz de darse cuenta de que la dosis de medicación debe ser una función del peso del paciente. La investigación reciente sugiere que los humanos razonan en gran medida a partir de estereotipos y de variaciones específicas de esos estereotipos. Nuestro sistema aún no ha demostrado esas habilidades.

Implementando sistemas de memoria asociativa

A corto plazo, los modelos de memoria asociativa complementarán bellamente a los modelos de IA. Los modelos asociativos se han estudiado largo rato, pero rara vez se los implementó (excepto para problemas muy pequeños) porque resulta muy caro hacerlos correr en computadoras digitales tradicionales. Una clase de implementación de memoria asociativa es el llamado modelo conexionista o de redes neuronales. Esos sistemas son descendientes directos de los modelos de redes neuronales de la década de 1950. En ellos, miles de unidades de procesamiento, cada una de ellas análoga a una neurona, se hallan interconectadas por nexos, cada uno análogo a una conexión sináptica entre neuronas. Cada nexo tiene un «peso» o fuerza de conexión. El conocimiento de un sistema se codifica en pesos de conexión y en el patrón de interconexiones del sistema. Algunas unidades sirven como unidades de entrada, algunas como unidades de salida y otras como «unidades ocultas» (están conectadas sólo a otras unidades y de este modo no pueden «verse» desde los canales de entrada o salida).

Estas redes exhiben tres características interesantes. La primera es el *aprendizaje*. Se han desarrollado diversos métodos que permiten que ese sistema, siempre que se le suministren insumos particulares, pueda ser enseñado a producir los eductos deseados. La segunda habilidad interesante es el *recuerdo asociativo*. Una vez entrenada para asociar un educto con determinado insumo, una red puede, ante una fracción del estímulo, producir el patrón completo del educto correspondiente. La tercera propiedad interesante es la *tolerancia a fallas*: la red continúa operando

incluso cuando se eliminan o se le dañan algunas unidades. En pocas palabras, los sistemas de computación conexionistas poseen muchas de las propiedades que se han asociado con los cerebros; estos sistemas difieren significativamente de las computadoras, que tradicionalmente han sido vistas como autómatas con mentes literales, capaces de hacer sólo aquello para lo cual han sido programados.[14]

Estas redes ya se pueden implementar eficientemente en hardware masivamente paralelo como el del sistema de la Máquina de Conexión o utilizando chips especiales. Cuando los sistemas de memoria asociativa se simulaban en computadoras digitales seriales comunes, las simulaciones fueron muy lentas; una computadora serial debe simular varias veces una por una todas las unidades computacionales y nexos para desarrollar un cálculo sencillo. Una máquina masivamente paralela puede asignar un pequeño procesador separado a cada una de las unidades en la memoria asociativa y de este modo operar mucho más rápidamente.

Stanfill y yo hemos estado explorando un método masivamente paralelo funcionalmente similar llamado razonamiento basado en la memoria (o en los recuerdos). En este tipo de razonamiento, se carga una Máquina de Conexión con una gran base de simulaciones. Cada situación en la base de datos contiene tanto un conjunto de atributos como una consecuencia. En una base de datos médica, por ejemplo, los atributos serían síntomas y características del paciente, y las consecuencias un diagnóstico o un tratamiento. Cada ítem en una base de datos se almacena en un procesador separado. Cuando se encuentra un nuevo ejemplo a clasificar, sus propiedades se transmiten a todos los procesadores que almacenan situaciones; cada uno de estos procesadores compara la situación que tiene almacenada con la del insumo y computa un puntaje de similitud. El sistema encuentra entonces las correspondencias más próximas al insumo y, siempre que sean lo suficientemente estrechas, utiliza las consecuencias de esos ítem coincidentes para clasificar el nuevo ejemplo.

Los sistemas de razonamiento basados en la memoria poseen asimismo diversas características oportunas. Son tolerantes a fallas; pueden generalizar bien ante ejemplos que nunca han sido vistos en su forma exacta con anterioridad; pueden medir el parecido entre los precedentes de un ejemplo actual, lo que puede

oficiar como medida del valor de confianza de la comparación. Si hay un caso de exacto parecido con un ejemplo previo, el sistema puede tomar una decisión con certidumbre. Es fácil enseñar a estos sistemas: simplemente se añaden más elementos a sus bases de datos.

La parte complicada de los sistemas de razonamiento basado en la memoria es la computación de la similitud. Para calcular el parecido entre cualquier ejemplo en la memoria y el patrón que debe clasificarse, cada ítem de memoria debe encontrar primero la distancia (o diferencia) entre cada uno de los valores de sus atributos y los valores de los atributos del patrón a clasificar. Estas distancias dependen a su vez de la distribución estadística de los valores de atributo y del grado de correlación entre cada valor de atributo y la consecuencia que se manifiesta simultáneamente con él. Se deben combinar todas las distancias para cada atributo para cada ítem de memoria para obtener la distancia total con respecto al ítem a clasificar. De esta forma, computar el puntaje de parecido involucra un montón de cálculo estadístico a través de todos los registros en la base de datos.[15]

¿Cuál es el papel de los sistemas de memoria asociativa en la inteligencia artificial tradicional? Pese a que pueden sustituir a los sistemas expertos bajo ciertas circunstancias, los sistemas de razonamiento conexionistas y los sistemas basados en la memoria se contemplan mejor como complementos de la IA tradicional que como sus sustitutos. En una modalidad muy útil, los sistemas de memoria asociativa se pueden utilizar para proponer o hipotetizar soluciones a problemas complejos, y los sistemas tradicionales de IA se pueden usar para verificar que las diferencias entre los problemas que están siendo atacados y los ejemplos de la base de datos sean de poca importancia. Si esas diferencias son importantes, los sistemas de memoria asociativa pueden proponer que se intenten otros subobjetivos. De esta forma, el proceso de memoria asociativa puede proporcionar un método heurístico sumamente poderoso para saltar a conclusiones, mientras que la IA tradicional puede utilizarse para verificar o desconfirmar tales conclusiones. Esos sistemas híbridos pueden ayudar a que los modelos de IA eviten los problemas de buscar en forma combinatoria en grandes espacios de soluciones. Dados los recursos computacionales requeridos, el núcleo de la fuerza computacional en un sistema de IA de este tipo probablemente resida en la porción de memoria asociativa.

A largo plazo, sin embargo, es todavía improbable que esos modelos proporcionen una explicación satisfactoria de las operaciones del pensamiento humano, aunque sospecho que llegarán más cerca que la IA. A mi entender, la mejor exposición sobre la arquitectura definitiva requerida es la «sociedad de la mente» de Marvin Minsky.[16] Minsky argumenta persuasivamente, utilizando un rango muy amplio de tipos de evidencia, que el cerebro y la mente están compuestos de un número muy grande de módulos organizados como una burocracia. Cada módulo o «demonio» en esa burocracia sólo tiene responsabilidades limitadas y un conocimiento muy limitado; los demonios vigilan constantemente en espera de sucesos que son de interés para ellos, y sólo actúan cuando ocurre uno de esos sucesos. Estos sucesos pueden ser externos (señalados por unidades sensoriales) o puramente internos (los resultados de otros demonios internos que han reconocido elementos de su propia órbita de interés). Las acciones de los demonios pueden ejercer influencia sobre otros demonios o activar efectores que les den así influencia sobre el mundo exterior. Se puede trazar una analogía simple entre la sociedad de la mente y los modelos de memoria asociativa: en el razonamiento basado en la memoria, cada ítem de la base de datos correspondería a un agente; en un modelo conexionista, cada unidad neuronal correspondería a un agente.

Razonamiento lógico

Creo que el razonamiento lógico no es el fundamento sobre el que se construye la cognición, sino una conducta emergente que resulta de observar un número suficiente de regularidades en el mundo. De esta manera, si una sociedad de agentes parecidos a demonios exhibe una conducta lógica, su conducta puede describirse mediante reglas, aunque el sistema no contenga reglas que gobiernen su operación. Opera en forma regular porque simula las regularidades del mundo.

Consideremos un bebé en desarrollo. En el modelo de la sociedad de la mente, el niño desarrolla primero un gran número de agencias independientes que codifican conocimiento sobre la conducta de tópicos específicos en el mundo físico: cuando se suelta un bloque, éste cae; cuando el niño llora, sus padres vienen

a atenderlo; cuando el niño toca una llama, siente dolor. Cada uno de estos ejemplos es manipulado inicialmente por una pequeña burocracia de agentes. Cada burocracia representa el recuerdo de algún suceso específico. Una agencia particular es responsable de un episodio, debido al «cableado» original del cerebro; poco después que una agencia se activa por primera vez, cambia sus pesos sinápticos de manera que cualquier nuevo suceso que active cualquier parte de la agencia ocasionará que toda la agencia sea reactivada. Cuando sucesos similares reactivan esas agencias, se construyen nuevas burocracias que codifican las similitudes y diferencias entre los nuevos y los viejos sucesos, a partir de agentes que previamente estaban sin usar, pero que estaban estrechamente conectados (y por ello, activados). Después de muchas de esas adiciones incrementales a la sociedad de agentes, un niño desarrolla con el tiempo agentes para las categorías abstractas y las reglas; las cortaduras, los pinchazos y las quemaduras causan dolor, y así otros agentes que se activen en esos casos quedan asociados al concepto de dolor. Con el tiempo, los conceptos de la conjunción constante del dolor con sus diversas causas se convierten en la especialidad de agentes «expertos» particulares, responsables ante determinadas regularidades del mundo. En última instancia, esos agentes devienen parte de la burocracia para el concepto mismo de causalidad. De este modo, los agentes llegan a razonar sobre categorías sumamente generales, ya no necesariamente enraizadas en forma directa en la experiencia, y pueden comprender relaciones causales abstractas. Tomemos el dolor en abstracto, por ejemplo: si uno quiebra una ley y es aprehendido, sabe que probablemente será castigado; si uno no cumple las promesas, comprende que otra gente puede enojarse y que se puede vengar y así sucesivamente.

En la superficie puede parecer que lo que se ha propuesto es reemplazar un solo sistema experto por muchos sistemas expertos, dispuestos en una jerarquía. Sin embargo, cada uno de los sistemas expertos es extremadamente simple, en el sentido de que «conoce» una sola cosa. Los expertos se hallan conectados a un sistema perceptual y entre sí de tal manera que se disparan sólo cuando se satisfacen realmente las condiciones respecto a las cuales son expertos.

Aunque ésta puede ser una descripción satisfactoria de la composición de la mente, todavía no es lo suficientemente precisa

como para servir para el diseño de un programa en muy gran escala que se pueda organizar para alcanzar inteligencia. Los programas que operan según el principio de la sociedad de la mente bien pueden ser el punto final de numerosas etapas en la evolución del diseño de los sistemas inteligentes. Creo que los híbridos de la memoria asociativa y de los programas tradicionales de IA para el razonamiento lógico representan la principal promesa a corto plazo para las aplicaciones de IA. Es posible que también demuestren ser modelos útiles de la cognición.

Los límites del hardware de computación tradicional

La sospecha de los investigadores respecto de que los modelos actuales de IA pueden no ser extensibles a sistemas con inteligencia de nivel humano no es la única fuerza que empuja al cambio de paradigmas hacia los modelos de computación masivamente paralelos. Las consideraciones económicas, que trascienden las preocupaciones de la IA, son otras. Las computadoras seriales actuales han comenzado a alcanzar los límites más allá de los cuales no podrán aumentar su velocidad a un costo razonable. Para que una computadora serial de un solo procesador opere con más rapidez de la que opera actualmente, su procesador debe ejecutar cada instrucción con mayor velocidad. Para acelerar el procesamiento, los fabricantes han introducido el uso de nuevos materiales, que permiten una acción más rápida. También han encogido los circuitos a tamaños cada vez más pequeños para acortar la longitud del paso de las señales, dado que las velocidades internas de comunicación, y por lo tanto la velocidad general de procesamiento, se hallan limitadas por la velocidad de la luz. Cuanto más pequeña es una computadora, más veloces son sus comunicaciones internas. Dado que cada componente genera calor, y dado que los chips densos producen más calor que los otros, los chips ultradensos de materiales exóticos requieren a menudo que se agreguen costosos y elaborados sistemas de enfriamiento. Todo esto significa que duplicar la potencia de una máquina serial habitualmente aumenta su costo por un factor mayor que dos, a veces mucho mayor.

En contraste, los diseños paralelos prometen la posibilidad de

duplicar la potencia simplemente duplicando el número de procesadores, posiblemente por menos de dos veces el costo, dado que muchos de los componentes del sistema (unidades de almacenamiento en disco, fuentes de potencia, control lógico, etcétera) son compartidos por todos los procesadores, sin importar cuántos sean. Por ejemplo, la Máquina de Conexión contiene 65.536 procesadores. Aun en su versión original, la Máquina de Conexión es poco cara en términos del monto en dólares que cuesta por cada unidad de computación; su costo en relación con su performance es de alrededor de la vigésima parte del costo de las supercomputadoras seriales. (La cifra de costo/performance es la operación estándar de costo en computación. La típica operación de computación estándar es o bien una suma con punto fijo o una multiplicación con punto flotante. La performance con punto fijo se mide en millones de instrucciones por segundo [MIPS]; la performance con punto flotante se mide en millones de ejecuciones con punto flotante por segundo [MFLOPS, pronunciado «megaflops»]; el costo/performance se mide en dólares por MIPS o dólares por MFLOPS.) Más aún, es probable que el costo de los procesadores altamente paralelos descienda en forma dramática. Al principio, cualquier chip es caro debido a la escasa producción (sólo una fracción de chips utilizables resultan de la producción inicial) y a la necesidad de amortizar la investigación, el diseño y los costos de desarrollo. El precio de los chips sigue una «curva de aprendizaje», una caída en costos que es una función del número de chips fabricados. La memoria es el ejemplo más adecuado: el costo por bit de almacenamiento en memoria ha caído por un factor de diez cada cinco años durante los últimos treinta, teniendo ahora un costo que es de un diezmillonésimo del precio de 1950, ¡un cienmillonésimo si ajustamos la inflación! Dado que los procesadores de una computadora masivamente paralela se producen en masa, igual que los chips de memoria, el costo de una cantidad determinada de fuerza de procesamiento para las máquinas paralelas debe caer con tanta rapidez como el costo de la memoria, es decir, muy rápidamente.

El costo de los sistemas de computadora involucra, por supuesto, tanto al hardware como al software. ¿Cómo ha de programarse una máquina con decenas de miles o quizá millones de procesadores? Claramente, los programadores humanos no pueden afrontar el tiempo o el dinero necesarios para escribir un

programa para cada procesador. Parece haber dos formas prácticas para programar esas máquinas. La primera, que es la que más se ha usado hasta ahora, es escribir un solo programa y hacer que cada procesador lo ejecute en sincronía, cada uno de ellos encargándose de su propia porción de datos. Es el método de «paralelismo a nivel de los datos». Una segunda forma es programar máquinas de aprendizaje que pueden volcar sus experiencias en códigos o datos diferentes para cada procesador.

La investigación en aprendizaje de máquina ha crecido en forma dramática en los últimos años. Los investigadores han identificado quizás una docena de métodos de aprendizaje diferentes.[17] Muchos de los sistemas de aprendizaje masivamente paralelo involucran los modelos conexionistas o de redes neuronales que mencionamos antes. Los sistemas conexionistas usualmente han sido entrenados con alguna forma de aprendizaje supervisado: un insumo y una salida deseada se presentan al sistema, el cual ajusta las fuerzas de conexión internas entre sus unidades similares a neuronas de modo de hacerlas coincidir estrechamente con la conducta de entrada-salida correspondiente. Dado un número suficientemente grande de intentos, generalmente del orden de las decenas de miles, esos sistemas son capaces de aprender a producir conductas definidas moderadamente complejas. Por ejemplo, después de comenzar con un estado inicial com-pletamente al azar y de ser entrenado repetidamente con una base de datos de 4.500 palabras de muestras de pronunciación, un sistema llamado NETtalk fue capaz de aprender a pronunciar nuevas palabras en inglés con una exactitud bastante apreciable.[18]

El problema central a resolver en la investigación sobre el aprendizaje conexionista y de la sociedad de la mente es el «problema de la asignación de crédito», el problema de asignar recompensas y castigos simples entre un vasto número de elementos de computación interconectados similares a las neuronas. Para mostrar la relevancia de este problema para los objetivos últimos de la IA, lo explicaré en términos del «cerebro» de un sistema robótico que, según esperamos, aprenderá de sus experiencias.

Supongamos un gran conjunto (quizá miles de millones) de elementos de procesamiento de tipo neuronal independiente, interconectados con muchos nexos por cada elemento. Algunos elementos están conectados a sensores, determinados a su vez

por el mundo exterior; otros están conectados a sistemas motores que pueden ejercer influencia sobre el mundo exterior a través de brazos, piernas o ruedas robóticas, que generan actos físicos, así como también mediante facilidades de producción de lenguaje, que generan «actos de habla». En cualquier momento dado, un subconjunto de esos elementos se halla activado; forman un complejo patrón de activación a través de toda la red. Poco tiempo después, los patrones de activación cambian debido a las influencias mutuas entre los elementos procesadores y los insumos sensoriales.

Algunos patrones de activación disparan acciones motoras. Aquí y allá se asignan al sistema recompensas o castigos. El problema de la asignación de crédito es éste: ¿qué elementos individuales dentro de una masa de quizá trillones de elementos deben alterarse sobre la base de estas recompensas y castigos de modo que el sistema aprenda a ejecutar más efectivamente, es decir, de modo que en el futuro se eviten las situaciones que conducen al castigo, para que el sistema se encuentre con mayor frecuencia en situaciones que acarrean recompensas?

El problema de la asignación de crédito tiene por lo menos dos aspectos. El más simple es el del problema de la asignación de crédito *estática*, en la que las recompensas y castigos ocurren inmediatamente después que las acciones que los acarrean. Estos sistemas reciben gratificación instantánea y retroalimentación negativa instantánea. El problema de la asignación de crédito estática se ha encontrado razonablemente tratable: las unidades que están activas se pueden examinar, y aquellas que lo están en la dirección correcta poseerán sus conexiones con los sistemas de acción reforzados, mientras que las que se han activado en forma inapropiada los tendrán debilitados. Si la recompensa o el castigo ocurren sustancialmente después del hecho, sin embargo, tenemos un problema de asignación de crédito *temporal*, que es significativamente más difícil. Para resolver este problema, un sistema debe guardar recuerdo de los estados anteriores por los que ha pasado y tener la capacidad de analizar y hacer juicios sobre cuáles de sus estados anteriores han sido responsables de las recompensas y castigos. El progreso en el problema de la asignación de crédito temporal ha sido promisorio, pero hay mucho por hacer antes que se lo pueda considerar resuelto.[19]

Según estimo, estos métodos de aprendizaje sólo serán viables

para producir módulos de un sistema inteligente mayor. Un sistema verdaderamente inteligente deberá contener muchos módulos. Parece muy improbable que la organización de todo un cerebro o una mente se pueda aprender automáticamente, comenzando con sistemas muy grandes y conectados al azar. Los bebés están sumamente organizados cuando nacen. Por ejemplo, no tiene que aprender a ver o a oír en ningún sentido que pudiéramos reconocer como aprendizaje. Sus sistemas auditivos y visuales ya parecen organizados para poder extraer unidades significativas (objetos, sucesos, sonidos, formas, etcétera). Elizabeth Spelke y sus investigadores asociados han encontrado que los bebés de dos meses son capaces de reconocer la coherencia de objetos y que muestran sorpresa cuando los objetos desaparecen o cuando parecen moverse unos a través de otros.[20] A esa edad, los bebés no pueden haber aprendido las propiedades de los objetos mediante la experiencia táctil. No sería sorprendente que esas habilidades estuvieran «precableadas» en el cerebro: los caballos y terneros recién nacidos pueden caminar, evitando tropezarse con los objetos, y encuentran la leche de su madre pocos minutos después de nacer. En todo caso, la necesidad de proporcionar una organización sensorial a priori a los sistemas inteligentes parece ineludible. ¿Sobre qué otras bases podríamos aprender desde la nada cuáles son las unidades significativas del mundo?[21]

El futuro de la inteligencia artificial

Cualquier extrapolación de las tendencias actuales nos fuerza a llegar a la conclusión de que tomará un largo tiempo lograr sistemas que sean tan inteligentes como los humanos. Sin embargo, la performance de las computadoras más rápidas parece destinada a incrementarse a un ritmo mucho mayor que el que ha sido el caso en los últimos treinta años, y las cifras de costo/performance de las computadoras en gran escala con seguridad caerán.

El efecto de un poder de procesamiento mucho mayor será sumamente significativo para la IA. Como afirmé antes, las máquinas actuales sólo tienen un cuatrillonésimo de la capacidad de computación que el cerebro humano. Como sea, es concebible

que dentro de unos veinticinco años podamos construir máquinas de potencia comparable por precios accesibles (para los propósitos de esta argumentación, pongamos que un precio accesible sean 20 millones de dólares, el costo de la supercomputadora actual más cara).

El sistema de la Máquina de Conexión, en la actualidad quizás el más rápido del mundo, puede desarrollar los tipos de computación que pensamos que usa el cerebro a unos 3.6×10^{12} bits por segundo, un factor de alrededor de veinte millones de veces demasiado lejos para poder equipararse a la potencia del cerebro (tal como la estimó Jack Schwartz en su artículo en esta edición de *Dædalus*). Se puede construir una Máquina de Conexión más poderosa simplemente conectando juntas varias de ellas. La máquina actual cuesta unos 4 millones, de modo que, dentro de nuestro presupuesto de 20 millones, se puede construir una máquina de unas cinco veces su potencia de cómputo, o unos 1.8×10^{13} bits por segundo). Esa máquina se quedaría corta en un factor de unos 4 millones. El objetivo declarado de la Iniciativa de Computación Estratégica de la DARPA (Defense Advanced Research Projects Agency) es alcanzar un incremento de mil veces el poder de cómputo en los próximos diez años, y hay buenas razones para creer que se alcanzará este objetivo. En particular, el sistema de la Máquina de Conexión alcanza sus ritmos de computación sin utilizar todavía materiales exóticos o una miniaturización extrema, factores que nos han permitido acelerar las computadoras tradicionales. Si se puede alcanzar una aceleración de mil veces cada diez años, podría construirse una computadora comparable al cerebro en potencia de procesamiento por 20 millones de dólares hacia el año 2012.

Utilizando las estimaciones de Schwartz, encontramos que la capacidad total de memoria del cerebro es de 4×10^{16} bytes. La Máquina de Conexión actual puede contener hasta dos gigabytes (2×10^{9} bytes). En el mundo de las computadoras de hoy, dos gigabytes de memoria se considera una cantidad grande, aunque éste es un factor que se queda corto en veinte millones, o un factor cuatro millones de veces corto para un sistema con cinco Máquinas de Conexión.

A los precios actuales, dos gigabytes de memoria cuestan alrededor de un millón de dólares, de modo que comprar suficiente memoria para equiparar la capacidad humana costaría unos 20

mil billones, más o menos diez veces la deuda nacional. Dado que su precio a largo plazo declina por un factor de diez cada cinco años, el costo de 4×10^{16} bytes de memoria estará en el rango de los 20 millones en treinta años, de modo que la época en que podemos esperar construir una computadora con un potencial equiparable al de la inteligencia humana estará en torno del año 2017. (Mucho antes de la fecha de 2017, sin embargo, los dispositivos de almacenamiento masivo —unidades de disco y otros soportes de almacenamiento— serán capaces de almacenar gran parte de este material a un precio accesible.) Como antes se sugirió, no obstante, construir el hardware puede ser la parte más fácil; la necesidad de desenredar los misterios de la estructura y el funcionamiento de la mente, de reunir tanto el conocimiento innato como el adquirido y de diseñar el software para todo el sistema requerirá probablemente unos cuantos años más allá del 2017. Una vez que tengamos una pieza de hardware con potencia a nivel del cerebro y una estructura a priori adecuada, ¡todavía puede llevar tanto como veinte años más alcanzar la competencia mental de un adulto! Es posible que se necesite más de uno de esos prolongados experimentos.

¿A qué podemos esperar que se parezca la inteligencia de una de esas máquinas poderosas? Casi con certeza, las máquinas parecerán extrañas en comparación con la gente. En ciertos aspectos opacarán la máxima performance humana, del mismo modo en que las calculadoras de bolsillo superan a los humanos en cálculos matemáticos. Las nuevas máquinas tendrán memoria perfecta de vastas cantidades de información, algo que no es posible para la gente. (Mientras los humanos tienen en apariencia vastas cantidades de recuerdos, somos más bien pobres para la memorización literal de palabras, imágenes, nombres y detalles de sucesos.) A menos que se lo programe deliberadamente, esas máquinas no tendrán un repertorio de emociones humanas reconocibles. Ni pueden tener motivación en ningún sentido humano ordinario. La motivación y el impulso parecen basarse en mecanismos innatos desarrollados durante eones de evolución para asegurar que tomamos decisiones preservativas para la especie —evitar el pánico, seguir comiendo y bebiendo, dormir suficientemente, reproducirse, cuidar de los jóvenes, actuar con altruismo (en especial hacia los parientes y amigos)— sin requerir que comprendamos que la razón real para realizar esas acciones

sea la preservación de la especie.[22] (Es muy posible, sin embargo, que resulte útil dotar a las máquinas capaces de resolver problemas y aprender con la capacidad de experimentar algunos análogos de la frustración, el placer de alcanzar un objetivo, la confusión y otras actitudes relacionadas con las emociones hacia los fenómenos emergentes, para que puedan generar abstracciones útiles para decidir cuándo abandonar una tarea, pedir consejo o renunciar.)

Los investigadores en IA pueden lograr la oportunidad de construir máquinas inteligentes a nivel humano sólo si encuentran la forma de llenar cantidades prodigiosas de memoria con material importante. Podrán hacerlo sólo si la IA puede producir sistemas sensoriales adecuados (para el oído, la visión, el tacto, la cinestesia, el olfato y el gusto). Con sistemas sensoriales, los sistemas de IA serán capaces por primera vez de aprender de la experiencia. Esa experiencia será al principio poco más que aprendizaje de memoria, es decir, almacenar registros de patrones sensoriales parcialmente digeridos vistos por el sistema. Sin embargo, como se afirmó antes, el almacenamiento de vastas cantidades de material relativamente literal puede ser una de las claves de la conducta inteligente. El potencial de la IA depende de la posibilidad de construir sistemas que ya no requieran programación en el mismo sentido en que hoy se requiere. Podremos superar entonces la tendencia que tiene el desarrollo de sistemas a hacerse lento debido a las dificultades del diseño del software.

También está la cuestión de la clase de «cuerpo» en que deberá incrustarse esa inteligencia para comprender verdaderamente, más que para simular la comprensión. ¿Debe cablearse la máquina para que tenga emociones, si es que ha de comprender nuestras reacciones emocionales humanas? Si una máquina fuera inmortal, ¿podría comprender nuestras reacciones al conocimiento de nuestra propia mortalidad? Las máquinas inteligentes pueden clonearse simplemente copiando su programa o su código interno sobre otras piezas idénticas de hardware. No hay un equivalente humano a una máquina que pueda tener la experiencia de ser una entidad unitaria durante un largo período y luego, en algún momento de su «vida», devenir de repente muchas entidades separadas, cada una con diferentes experiencias. Qué clase de inteligencia exactamente sea ésta es por lo tanto una cuestión abierta.[23]

Resumen

Estamos cerca de un importante acontecimiento en la historia de la vida en la tierra, el punto en el que podemos construir máquinas con el potencial de exhibir una inteligencia comparable a la nuestra. Parece seguro que podremos construir hardware que equipare la potencia de cómputo humana por un precio accesible dentro de los próximos treinta años o una cosa así. Sin duda, ese hardware tendrá profundas consecuencias para la industria, la defensa, el gobierno, las artes y nuestras imágenes de nosotros mismos.

Tener hardware con una potencia a nivel de la del cerebro, sin embargo, no lleva a sistemas inteligentes de nivel humano, dado que la arquitectura y los programas para tales sistemas también presentan obstáculos sin precedentes. Es difícil extrapolar los efectos futuros a partir del ritmo de progreso que se ha venido dando hasta ahora. El progreso ha sido muy lento, en parte porque los modelos computacionales que se han venido usando han sido inapropiados para la tarea. Esta impropiedad se aplica más críticamente al problema del aprendizaje. Sin aprendizaje, los sistemas serán construidos a mano. No sabemos con qué exactitud debemos equiparar los detalles del cerebro humano para nutrir un aprendizaje y una performance apropiados. Con las arquitecturas correctas, es probable que el progreso se acelere, tanto en lo que respecta a construir el hardware adecuado como en lo que hace a programarlo. Creo que la construcción de máquinas verdaderamente inteligentes es lo suficientemente probable para que resulte justificado comenzar a estudiar y planear su política ya mismo. De ese modo podremos maximizar sus beneficios y minimizar sus efectos negativos sobre la sociedad.

Notas

[1] Frederick P. Brooks, *The Mythical Man-Month: Essays on Software Engineering* (Reading, Mass.: Addison-Wesley, 1974).

[2] Allen Newell y Herbert Simon, *Human Problem Solving* (Engelwood Cliffs, N.J.: Prentice-Hall, 1972).

[3] Los operadores booleanos Y, O y NO constituyen un conjunto universal. NY

(NO Y) es también universal por sí mismo, igual que NO-O. Para una derivación de este resultado véase Marvin L. Minsky, *Computation: Finite and Infinite Machines* (Cambridge: MIT Press, 1967).

[4] Herbert A. Simon, *The Sciences of the Artificial* (Cambridge: MIT Press, 1965).

[5] John Backus, «Can Programming be Liberated from the von Neumann Style? A Functional Style and its Algebra of Programs», *Communications of the ACM* 21 (8) (agosto de 1978):613-41.

[6] George A. Miller, Eugene Galanter y Karl Pribram, *Plans and the Structure of Behavior* (Nueva York: Holt, Rinehart y Winston, 1954).

[7] John McCarthy, «Epistemological Problems in Artificial Intelligence», en *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (Los Altos, Calif.: Morgan-Kaufmann, agosto de 1977), 1038-44.

[8] Los ejemplos son de Marvin Minsky, comunicación personal.

[9] John McCarthy, «Circumscription - A Form of Nonmonotonic Reasoning», *Artificial Intelligence* 13 (1) (1980):27-39; y Drew V. McDermott y Jon Doyle, «Nonmonotonic Logic I», *Artificial Intelligence* 13 (1) (1980):41-72.

[10] Steve Hanks y Drew V. McDermott, «Default Reasoning, Nonmonotonic Logics, and the Frame Problem», en *Proceedings of the Fifth National Conference on Artificial Intelligence* (Los Altos, Calif.: Morgan-Kaufmann, agosto de 1986), 328-33.

[11] Craig Stanfill y David L. Waltz, «Toward Memory-Based Reasoning», *Communications of the ACM* 29 (12) (diciembre de 1986):1213-28.

[12] W. Daniel Hillis, *The Connection Machine* (Cambridge: MIT Press, 1986).

[13] David E. Rumelhart, James L. McClelland y el PDP Research Group, editores, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volúmenes 1 y 2 (Cambridge: MIT Press, 1986).

[14] Para un tratamiento más extendido de estas cuestiones, véase Rumelhart y McClelland, *Parallel Distributed Processing* y la edición especial sobre modelos conexionistas de *Cognitive Science* 9 (1) (1985).

[15] Para más detalles, véase Stanfill y Waltz, «Toward Memory-Based Reasoning».

[16] Marvin L. Minsky, *The Society of Mind* (Nueva York: Simon y Schuster, 1986) [Traducción española: *La Sociedad de la Mente*, Buenos Aires, Galápago, 1986]; *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence*, de A. Harry Klopf (Washington, D.C.: Hemisphere, 1982), presenta una teoría neuronal compatible.

[17] También hay una literatura bastante abundante sobre el aprendizaje y la adquisición del conocimiento. Se basa en los paradigmas de la búsqueda heurística y del sistema físico de símbolos. Hablando en general, estos algoritmos de aprendizaje caen dentro de tres categorías. El primer tipo es estadístico y utiliza un gran número de procesadores para encontrar patrones o regularidades en bases de datos constituidas por ejemplos (en diagnóstico médico, en predicción meteorológica o en toma de decisiones, digamos). Tres sistemas que caen dentro de esta categoría son el sistema ID3 de Ross Quinlan y los sistemas construidos por Ryszard Michalski, ambos descritos en Ryszard S. Michalski, Jaime Carbonell y Thomas Mitchell, editores, *Machine Learning: An Artificial Intelligence Approach* (Los Altos, Calif.: Tioga Publishing Company, 1983), y el «sistema de razonamiento basado en la memoria» de Craig Stanfill y David Waltz (véase Stanfill y Waltz, «Toward Memory-Based Reasoning»). Un segundo tipo de algoritmo de aprendizaje utiliza «reglas de producción» (a veces llamadas «reglas si-entonces») y aprende enriqueciendo y modificando un conjunto existente de reglas. Las reglas se

cambian proporcionando «experiencia», la que puede incluir «recompensas y castigos». También se puede enseñar a esos sistemas dándoles ejemplos correctos de los que ellos pueden aprender reglas memorizándolos. Dos sistemas de esta clase son los algoritmos genéticos de John Holland (véase John H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence* (Ann Arbor: University of Michigan Press, 1975) y el sistema SOAR de Allen Newell y Paul Rosenbloom (véase John E. Laird, Paul S. Rosenbloom y Allen Newell, «Chunking and SOAR: The Anatomy of a General Learning Mechanism», *Machine Learning* 1 [1] [1986]:11-46). La tercera rama es la del «aprendizaje basado en la explicación (véase Gerald F. DeJong y Raymond A. Mooney, «Explanation-Based Learning: An Alternative View», *Machine Learning* 1 [2] [abril de 1986]:145-76). Un sistema de aprendizaje basado en la explicación intenta construir estructuras causales, o «schemata», como explicaciones de nuevos fenómenos y como elementos para construir schemata nuevos. Uno de los métodos de aprendizaje más exitosos es la pieza central de *Parallel Distributed Processing* de McClelland y Rumelhart. Otros sistemas incluyen los métodos de Stephen Grossberg, de Andrew Barto y de Geoffrey Hinton. Véase Stephen Grossberg, «Competitive Learning: From Interactive Activation to Adaptive Resonance», *Cognitive Science* 11 (1) (enero-marzo de 1987):23-64; Andrew G. Barto, «Learning by Statistical Cooperation of Self-Interested Neuron-like Computing Elements», *Human Neurobiology* 4 (1985):229-56; y Geoffrey Hinton, «The Boltzmann Machine», en Geoffrey Hinton y John A. Anderson, editores, *Parallel Models of Associative Memory* (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1981).

[18] Terrence J. Sejnowski y Charles R. Rosenberg, «NETtalk: A Parallel Network that Learns to Read Aloud», Technical Report JHU/EECS-86-01 (Baltimore, Md.: Johns Hopkins University, Electrical Engineering and Computer Science, 1986).

[19] Ronald J. Williams, «Reinforcement-Learning Connectionist Systems: A Progress Report» (manuscrito inédito, College of Computer Science, Northeastern University, noviembre de 1986).

[20] Elizabeth Spelke, «Perceptual Knowledge of Objects in Infancy», en Jacques Mehler, Edward C. T. Walker y Merrill Garrett, editores, *Perspectives on Mental Representation: Experimental and Theoretical Studies of Cognitive Processes and Capacities* (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1962).

[21] En la *Crítica de la Razón Pura* Immanuel Kant argumenta esencialmente esto: que «las formas innatas de la percepción humana y las categorías innatas de la comprensión humana imponen un orden invariante en el caos inicial de la cruda experiencia sensorial». Citado por Paul M. Churchland en *Matter and Consciousness* (Cambridge: MIT Press, 1984), 84.

[22] Véase Isaac Asimov, *I, Robot* (Nueva York: The New American Library of World Literature, 1950), 6, para una exploración temprana sobre la necesidad de esta clase de ética para robots, encarnada en tres leyes de la robótica:

1. Un robot no puede herir a un ser humano ni, mediante la inacción, permitir que un ser humano resulte herido; 2. Un robot debe obedecer las órdenes que les dan los seres humanos, excepto cuando esas órdenes entren en conflicto con la Primera Ley; 3. Un robot debe proteger su propia existencia en la medida en que esa protección no entre en conflicto con la Primera o la Segunda Ley.

Irónico, por cierto, porque los primeros robots móviles de gran inteligencia probablemente se incorporarán a tanques y aviones de combate.

[23] Para una discusión extendida e interesante de cuestiones relacionadas con ésta, véase *Mind Children* de Hans Moravec (Cambridge: Harvard University Press, 1988).

10
Haciendo que las máquinas (y la inteligencia artificial) vean

Anya Hurlbert y Tomaso Poggio

La visión es más que un sentido; es una inteligencia. Imagínese viendo televisión: de una luz parpadeante en una pantalla bidimensional, usted crea un mundo tridimensional de gente, lugares y cosas. Los procesos mentales que llevan desde el patrón de luz en su retina a una imagen interna del mundo son tan «inteligentes» como los análisis e interpretaciones que llevan a un médico de los síntomas hasta el diagnóstico. Pero aunque nos maravillemos ante el cerebro humano, somos más propensos a apreciar los poderes deductivos de un lógico que la habilidad de la persona promedio para reconocer un rostro.

Nosotros los humanos somos en gran medida animales visuales. Casi el cincuenta por ciento de nuestro neocórtex —la parte del cerebro que llegó a lo último en la evolución y que por ello es mayormente característica de los primates— se halla dedicada a la visión. La visión es tan integral para nuestra comprensión del mundo que la habilidad para «ver» significa no sólo decodificar señales de luz, sino también captar la fuerza de una argumentación verbal. ¿Qué nos impide entonces llamar a la visión inteligencia?

Esta cuestión posee especial relevancia para los investigadores en inteligencia artificial (IA). El objetivo declarado de la IA es el de

Anya C. Hurlbert. Estudiante de M. D. y Ph. D. en la Escuela de Medicina de Harvard, el programa de ciencias de la salud y tecnología del Instituto Teconológico de Massachusetts y el departamento de ciencias cerebrales y cognitivas del MIT.
Tomaso Poggio. Profesor del Departamento de ciencias cerebrales y cognitivas del MIT. Dirige investigaciones sobre visión de computadoras en el laboratorio de inteligencia artificial del MIT, donde es codirector del Centro para el Procesamiento de Información Biológica.

recrear la inteligencia en máquinas y, al mismo tiempo, comprender qué es la inteligencia. Tradicionalmente, la investigación en IA ha atacado asuntos tan vocingleramente avanzados como el razonamiento, la resolución de problemas y el lenguaje. Un supuesto indiscutido subyacente a la investigación en IA es que el objetivo de la inteligencia es enriquecer nuestra interacción con, y mejorar nuestro control sobre el mundo exterior. Una inteligencia sin cuerpo, todo lo eficaz que sea para resolver problemas matemáticos, no puede alcanzar este objetivo si es incapaz de sentir o afectar el mundo. La robótica, el estudio de la forma de unir la percepción con la acción, es por consiguiente un complemento crucial de la IA.

La investigación en robótica se segmenta naturalmente en dos esfuerzos principales: la visión de máquina y el movimiento del robot. El objetivo de la investigación sobre la visión de máquina es el de construir máquinas que puedan ver y, al mismo tiempo, comprender la visión. De la misma manera, la investigación sobre el movimiento robótico procura no sólo construir robots que puedan manipular cosas y trasladarse, sino también comprender el control motor. La tentación obvia de la investigación en IA es asignarle a la visión de máquina la tarea de suministrar los insumos a una máquina inteligente y asignarle al movimiento robótico la tarea de ejecutar sus salidas. En sus comienzos, la investigación en IA hizo exactamente eso, y al hacerlo excluyó tanto a la visión como al control motor del ámbito de la inteligencia.

La razón de este exilio temprano de la visión de máquina fue en parte la razón misma de que diéramos la visión por sentada y subestimáramos sus fuerzas cotidianas: ver parece ser fácil e inmediato. Pero el progreso en la visión de máquina en los últimos veinte años ha desbaratado esa ilusión y ha revelado una humillante ironía. La visión no sólo es «inteligente», sino más difícil de entender o de recrear que el más sofisticado razonamiento matemático. De hecho, la visión plantea problemas tan difíciles que la IA actual se halla mucho más cerca de desarrollar sistemas que pueden servir como doctores o abogados que de construir robots que puedan reemplazar a los jardineros o los cocineros.

La promesa de la visión

La visión de máquina es en la actualidad integral a todo el esfuerzo de la IA, no sólo por su poder, complejidad y cabal utilidad,

sino por un mensaje saliente que porta su estrategia ante el estudio de la inteligencia. Recientemente, la IA tradicional (el término que usamos para la investigación en IA que excluye la visión de máquina y la robótica) ha sido desafiada por un viejo cargo bajo un nuevo nombre, el «conexionismo». El desafío proviene de gestálticos modernos que creen que la disección lógica que la IA practica sobre la inteligencia no puede revelar nunca la verdadera estructura o capacidad de las fuerzas más profundas de la mente.

La disputa apunta a una tradición que ha tomado el núcleo de la hipótesis del sistema físico de símbolos de Newell y Simon. La hipótesis afirma que «un sistema físico de símbolos posee los medios necesarios y suficientes para la acción general inteligente».[1] Un sistema físico de símbolos es una «máquina», nacida o construida en el mundo real, que trata con símbolos físicos, «patrones físicos» que designan cosas en el mundo real. El sistema debe ser capaz de construir nuevos símbolos a partir de los viejos. A su vez, los nuevos símbolos deben instruir el sistema para crear, modificar o destruir otros símbolos. En última instancia el sistema debe ser capaz de influir sobre las cosas reales a través de sus símbolos de ellas. La hipótesis reivindica el estudio de lo que las computadoras pueden hacer: las computadoras son sistemas físicos de símbolos, y por lo tanto son inteligentes.

Para la IA tradicional, estudiar lo que las computadoras pueden hacer significa programarlas para que hagan cosas. La rebelión contra la IA tradicional es, en parte, una rebelión contra la forma en que programa. La estrategia de la IA ha consistido en seccionar los problemas en sus tareas discretas más pequeñas y atacar cada problema por vez, siguiendo reglas explícitas que podan un árbol de soluciones ramificadas. Los sistemas expertos, que siguen reglas hechas a la medida para derivar inferencias a partir de una base de datos hecha a la medida, constituyen el fruto vendible esgrimido por esta estrategia. MYCIN, un sistema basado en el conocimiento para el diagnóstico médico, razona a partir de síntomas hacia un diagnóstico diferencial siguiendo reglas de tipo «si-entonces» almacenadas en su base de conocimientos. Orientado por los resultados de pruebas de laboratorio para buscar a través de su base de datos todos los patógenos posibles, MYCIN tomará su instrucción siguiente de un enunciado «si-entonces» apropiado: «Si encuentras un bicho, busca entonces una droga». Aunque muchos frutos no están maduros (los médicos todavía están por emplear cualquier sistema

experto médico como algo más que una biblioteca de referencia computadorizada), los sistemas expertos en unas pocas áreas especializadas como la configuración de computadoras y la reparación en telecomunicaciones han demostrado ser útiles.

El lado arriesgado de la hipótesis dice que, como los seres humanos son inteligentes, ellos son sistemas físicos de símbolos. Parte del nuevo enojo en contra de la IA tradicional es en contra del hecho de considerar al hombre un sistema físico de símbolos; la catexis se concentra en la palabra *símbolo*. En sentido estricto, un símbolo físico puede ser texto en una pantalla, un conjunto de conmutadores electrónicos, una masa de neuronas o una corriente eléctrica a través de una membrana de células. Pero la IA tradicional ha tomado una definición aún más estricta: los símbolos son términos abstractos que denotan cosas conocidas y obedecen a una gramática circunscripta. Como los símbolos lógicos en las matemáticas o las palabras en el idioma lituano, los símbolos deben comportarse de acuerdo con reglas lógicas. Newell y Simon enumeraron el lenguaje de computadora LISP como un ejemplo de un sistema físico de símbolos, aunque la IA tradicional parece haberlo considerado el único. Para la IA tradicional, diseñar inteligencia ha llegado a ser equivalente a manipular los símbolos de lenguajes como el LISP para crear sistemas cuyos engranajes se mueven de acuerdo con las reglas de la lógica.

Los autores de la hipótesis del sistema físico de símbolos lo consideran una bendición para la ciencia computacional. Pero algunos consideran la bendición como una maldición sobre la inteligencia humana, si todo lo que los sistemas físicos de símbolos pueden hacer es manejar comandos de computadora. El muy afamado General Problem Solver (GPS), uno de los primeros programas que buscaron mecanismos universales para resolver enigmas, se derivó escuchando a la gente resolver problemas en voz alta. ¿Cómo podría incluso el mejor GPS, preguntan los ofendidos humanos, conducir un automóvil? Un conductor experimentado no aplica conscientemente reglas tales como «cuando se maneja a más de treinta millas por hora, poner la tercera marcha», o «cuando se cambia de velocidad, apretar el pedal del embrague», o «cuando nos aproximamos a una casilla de peaje, soltar el acelerador». Simplemente conduce, ejecutando las acciones necesarias en forma automática. O consideremos un mecánico que busca una llave para aflojar una tuerca. Decide cuál es la

herramienta apropiada para el trabajo mediante una rápida intuición, y no mediante una búsqueda sistemática a través de los implementos disponibles. Ningún sistema experto, argumentan los oponentes a la IA, puede maniobrar piezas de ajedrez con tanta inteligencia y a tanta velocidad como un maestro de ajedrez, a menos que sea infinitamente grande e infinitamente veloz.

La queja en contra del modelo de la inteligencia de la IA tradicional cristaliza en el conexionismo. Los conexionistas alegan que los rasgos característicos supremos de la inteligencia humana son, entre otros, el pensamiento asociativo y la capacidad para aprender y generalizar a partir de ejemplos. Afirman que esos rasgos no son captados por los procedimientos de búsqueda serial y las estructuras dendriformes de los sistemas expertos de la IA. En lugar de eso, la inteligencia surgirá sólo a partir de un hardware especial que reproduzca el paralelismo masivo del cerebro humano, en el cual un número inmenso de células interconectadas atacan diferentes partes de una misma tarea al mismo tiempo.

Lo que debe recordarse antes de sucumbir a las argumentaciones en contra de la estrategia tradicional de la IA es dónde y por qué *ha tenido* éxito. Las mismas cosas que la IA, en sus comienzos, tan anhelosamente denominó inteligentes —el razonamiento matemático, la comprensión del lenguaje, la lógica abstracta— son las cosas que los sistemas expertos hacen mejor. Nos atreveríamos a afirmar que las cosas que consideramos más desafiantes para la mente son simplemente las cosas de las que somos más conscientes, porque son las que hemos aprendido último en la evolución y consecuentemente las que hacemos menos bien. Los sistemas expertos tienen buena chance de sobrepasarnos en esas tareas conscientemente difíciles y no fallarán prematuramente por la finitud del progreso tecnológico. La debilidad intrínseca de los sistemas expertos médicos no radica en su actual incapacidad para abarcar el enorme dominio del conocimiento médico —con el tiempo, probablemente lo abrcarán—sino en su incapacidad para reproducir el arte de la medicina. Aunque aun un sistema experto actual probablemente supere de un interno falto de sueño en evocar una lista completa de pruebas prescritas de laboratorio, el interno siempre aventaja a la computadora en percibir que la desdicha es la causa de que el paciente haya perdido el apetito.

Las lagunas de la intuición y la intuición instantánea en un extremo, las habilidades perceptuales ordinarias como el reconocimiento del habla en el otro: éstas son las fuerzas de la mente que

a la IA tradicional más le cuesta modelizar. Son las actividades mentales de las que somos menos conscientes y más capaces. La evolución ha pasado milenios perfeccionando esos talentos inconscientes, y es más que lógico sugerir que para reproducirlos hay que desarrollar otras tácticas que las usadas exclusivamente (y en forma más bien pobre) por nuestra mente más consciente.

La visión es tal vez la más inteligente de las maquinaciones mentales escondidas a la conciencia. Sin embargo, los métodos para estudiar la inteligencia que ha desarrollado la visión de máquina no son recónditos ni mágicos, y no brillan ni en lógica ni en intuición. En la visión de máquina, lo mejor de la IA tradicional se encuentra con lo mejor del conexionismo para construir una ciencia que se erige aparte de ambas.

La visión de máquina ha destilado la filosofía que la orienta de las mismas fuentes que la IA tradicional. En el núcleo de la hipótesis del sistema físico de símbolos está la idea de que los símbolos deben ser cosas arbitrarias, independientes de la máquina subyacente y sin significado hasta que se los hace tenerlo. La visión de máquina convirtió esa idea en lo que llamamos su dogma central: la inteligencia puede estudiarse como un sistema abstracto de procesamiento de información, independientemente de la maquinaria en la que se ejecuta. La visión de máquina ha seguido el dogma a lo largo de un único sendero, la estrategia computacional. La estrategia computacional describe exactamente la información que recibe un sistema y la información que éste saca, y busca una forma de computación que pueda transformar el insumo en salida. Para los sistemas visuales naturales o artificiales, cualquier computación de ese tipo se halla coaccionada necesariamente por las propiedades del entorno, el ojo y la luz que viaja entre ellos. La clave para la computación correcta radica en descubrir y respetar esas coacciones. La visión de máquina ha convertido la búsqueda de las coacciones en una ciencia del mundo natural.

El problema de la visión

Para comprender la fuerza de la estrategia de la visión de máquina, se debe apreciar primero la dificultad de los problemas que ataca. Un estimado fragmento de anécdota apócrifa sobre Marvin Minsky, uno de los padres fundadores de la IA, ilustra la

dificultad de esa apreciación misma. Hace unos veinte años, Minsky asignó a un estudiante graduado un problema aparentemente tratable en un proyecto de verano: conectar una cámara a una computadora y hacer que la computadora describiera lo que veía. El proyecto de verano se expandió en una industria de investigación, y a despecho del enorme progreso de la visión de máquina, ese problema aún no ha sido resuelto.

Gran parte de ese progreso ha resultado de una penosa investigación sobre la forma de plantear el problema: ¿qué es lo que la visión hace? La sencilla respuesta es que la visión transforma señales de luz en representaciones internas de las cosas que las transmiten. La visión humana comienza con un patrón de luz bidimensional (una imagen) en cada retina y finaliza con una descripción de objetos tridimensionales en términos de su forma, color, textura, tamaño, distancia y movimiento. El primer obstáculo en la visión es la imagen retiniana misma: contiene una cantidad enorme, casi inimaginable de información. En la retina hay una matriz de más de cien millones de fotorreceptores. La lente del ojo focaliza la luz en la retina de tal manera que el mundo tridimensional se aplana y se mapea directamente en el mosaico fotorreceptor, y cada fotorreceptor corresponde a un punto particular en el campo visual. La cantidad de luz que cae en un solo fotorreceptor está determinada por la cantidad de luz reflejada por el objeto que ocupa la posición correspondiente en el campo visual. A su vez, la cantidad de luz que un objeto refleja depende de la cantidad de luz que cae sobre él (que depende, por ejemplo, de la distancia a que está de una lámpara, o si se encuentra a la sombra de otro objeto) y de la sustancia de que está hecho (metal brillante, terciopelo negro mate, gasa transparente, piel vegetal lustrosa), entre otras cosas.

Si por cada fotón que captura cada fotorreceptor depositara un grano oscuro de plata en el papel fotográfico detrás del ojo, podríamos descortezar instantáneas del mundo a partir de nuestras cámaras Polaroid integradas. Pero la retina no es tan vinculante, y la imagen que envía al cerebro es abstrusa. En cualquier instante, la imagen es una matriz de señales electroquímicas, en la que el tamaño de cada señal es proporcional a la cantidad de luz que golpea el fotorreceptor que la conduce. En cada segundo, el cerebro debe procesar unas cien de estas imágenes, a medida que el ojo pasea sobre un mundo siempre cambiante. De este modo, lejos de registrar fotografías estáticas, la retina transmite una corriente de información visual dinámica.

En la visión de máquina la imagen retiniana se traduce en una matriz bidimensional de pixels. Cada pixel es una diminuta subdivisión de la imagen y contiene un número que representa el tamaño de la luz transmitida por un sensor individual de luz (o, lo que es equivalente, la intensidad de la luz que golpea el sensor). El trabajo de la máquina consiste en ejecutar las manipulaciones matemáticas sobre la matriz de números para convertirla en matrices más expresivas: por ejemplo, matrices que explícitamente codifican las distancias entre los objetos y la cámara, o matrices que asignan un color a cada material diferente.

Una imagen típica de visión de máquina puede estar compuesta de un millón de pixels. Cada pixel contiene un número de ocho bits. La cantidad total de información, aunque mucho menor que la de una imagen retiniana humana, es todavía del vertiginoso orden de los ocho millones de bits. Multiplíquese ese número por el número de imágenes por segundo que debe enviar una cámara para poder imitar a un ojo humano, y el ritmo de transmisión de información se elevará por lo menos a unos cuantos cientos de millones de bits por segundo. De este modo, aun la operación matemática más simple que la máquina ejecuta sobre el flujo de imágenes requiere miles de millones de multiplicaciones y sumas por segundo. Un millón o más de computadoras personales trabajando juntas a duras penas podrían realizar el trabajo.

Irónicamente, el problema real en la visión es que toda la información que hay en una imagen nunca es suficiente. Se pierde demasiada información en la proyección del mundo tridimensional en una superficie de dos dimensiones, tornando muy ambiguos los valores de los pixels en una matriz muy grande. Considérese un típico error de una fotógrafa principiante: colocar a su sujeto delante de un poste telefónico para capturar el verde que hay a cada lado; pero en la foto el poste parece perforar la cabeza del sujeto. En la proyección de la escena tridimensional en la película bidimensional, se ha perdido información crucial sobre la profundidad. El tamaño del poste telefónico es un indicio que puede interpretarse por lo menos de dos maneras: ya sea como un poste telefónico grueso muy lejos, o como una vara fina que sobresale de la cabeza del sujeto. La ambigüedad de la profundidad es la más obvia; subsiste otra ambigüedad en la interpretación del brillo y la oscuridad. Si los valores de intensidad en un conglomerado de pixels es mucho más alto que en el conglomera-

do vecino, la grieta entre ambos pudo haber sido causada de muchas maneras. Quizá cayera una sombra sobre un trozo de papel, creando la ilusión de un borde entre el papel claro y el oscuro, o quizá cerca de una hoja negra hay una hoja de papel blanco. Los números en sí no lo dicen.

La visión de máquina formula este problema como sigue: dada una matriz bidimensional de valores de intensidad, encontrar la disposición tridimensional de objetos y superficies que la produjo. Tal como está formulado, el problema parece impenetrable. Los rasgos que hay que recuperar —los colores, texturas y relaciones espaciales entre objetos, la posición y el color de la fuente de luz— están enredados sin esperanza en una matriz de números. Pero la visión de máquina ha hecho un largo camino desde que el estudiante de Minsky comenzó a abordar el problema. En los últimos quince años ha surgido una idea de la estructura de la visión que nos permite desarmar el problema en secciones independientes y controlables. Primero que nada, la visión debe dividirse netamente en al menos dos etapas: la *visión inicial* (que determina dónde están las cosas) y la *visión de alto nivel* (que determina qué son). En segundo lugar, la visión inicial misma se debe estudiar como un conjunto de módulos visuales separados, cada uno de los cuales extrae de la imagen un tipo distinto de información visual. Al encontrar la perspectiva correcta, la visión de máquina ha ido más allá de los límites de la IA tradicional y ha construido una ciencia sólida en sí misma: la ciencia de la óptica inversa.

Una mirada a la visión de máquina

El sentido de la estructura de la visión no surgió de inmediato. Los primeros esfuerzos de los científicos de la visión intentaron, en efecto, resolver el problema por las buenas o por las malas: se explotaba cualquier táctica que pudiera resolver la ambigüedad de la imagen, sin importar lo restringida que fuera la tarea en cuestión. Esta estrategia produjo *sistemas expertos* en visión: programas de interpretación de imágenes que consultaban un almacenamiento de conocimiento hecho a la medida para alimentar la aplicación de reglas hechas a la medida. Un ejemplo exagerado de uno de esos programas aborda la tarea de localizar el teléfono en una foto de un escritorio de oficina atestado. Entre

los supuestos que formula el programa para acotar su búsqueda, está que el teléfono es negro y que se encuentra a una altura y a una distancia fijas de la cámara que tomó la foto. La búsqueda que ejecuta el programa es relativamente fácil: barre la hilera apropiada de pixels en busca de un conglomerado de pixels de bajo valor que significan un objeto oscuro y luego verifica que el tamaño del conglomerado coincida con la distancia a que debería estar el teléfono cuando se lo ve desde la distancia especificada. Si los supuestos demuestran ser verdaderos, el programa ejecuta bien. Pero fracasa miserablemente si el teléfono es blanco o si se lo ha movido del escritorio al suelo.

Los soluciones *ad hoc* (o «hacks») generadas por esta estrategia dejan poco espacio para el desarrollo o la aplicación de principios científicos generales. En vez de delinear con claridad los pasos que debe seguir un sistema visual para ir de la imagen a la representación de los objetos, los primeros programas de visión mezclaban los niveles en desorden, aplicando decisiones de alto nivel para tratar con información de bajo nivel registrada por los valores de la matriz de pixels. Como muchos otros sistemas expertos incipientes en otros dominios, los primeros sistemas de visión de máquina sólo podían funcionar en ambientes restringidos y artificiales. Sus técnicas para afrontar minimundos circunscriptos no pudieron generalizarse para tratar con los entornos impredecibles en los que los humanos y los animales más primitivos explotan la visión con tanta eficiencia.

La visión temprana

La debilidad de las primeras estocadas al problema de la visión provocó un movimiento hacia la visión temprana. Lo que los pioneros del movimiento advirtieron fue que, sin una teoría de la comprensión de la imagen suficientemente general como para instruir la interpretación de cualquier imagen, la visión de máquina estaría condenada a la perpetuación infinita de *hacks*, cada uno mejor concebido que el anterior, pero ninguno capaz de tocar la flexibilidad para todo propósito del sistema visual humano. El camino hacia esa teoría general pasaba por una ciencia del mundo, no por una ciencia de la mente: un análisis en profundidad de la física de la interacción entre luz, ojo y objeto.

Los esfuerzos concentrados en los últimos quince años han producido un esquema de la visión temprana cuyo primer objetivo es el de construir un mapa de la escena que registra, para cada superficie opaca en la imagen, su distancia y orientación relativa al observador. Es decir, la visión temprana busca transformar la matriz inicial de valores de pixels en otra matriz de números que explícitamente agrupa partes de la imagen que son partes de una cosa y que dice dónde está cada cosa, en relación con el observador. Las «cosas» de la visión temprana son sólo partes de objetos, las caras visibles y los lados de totalidades mayores y todavía desconocidas. Este mapa, llamado «esquema $2^{1/2}$D» por David Marr, proporciona el resorte gracias al cual pueden perseguirse los siguientes objetivos de la visión temprana: asignar forma, color, textura, velocidad y dirección de movimiento a cada cosa en la imagen. Se pudieron dibujar entonces mapas separados, cada uno de los cuales registraba un tipo distinto de información visual, y cada uno de esos mapas se superpuso al registro del esquema $2^{1/2}$D. Alineando así el mapa de color con el esquema $2^{1/2}$D, por ejemplo, un observador total imaginario podría reaccionar a la distancia, orientación y color de cada superficie individual de la imagen.

La visión temprana debe cumplimentar dos tareas en el proyecto de computar las propiedades visuales individuales de cada superficie sólida en la imagen: 1) comprimir la abundante información de la imagen a sus rasgos más importantes, y 2) reducir la ambigüedad de esa información.

Detección de bordes

El primer paso para hacer el primer mapa es delinear los bordes entre las diferentes regiones de la imagen. La *detección de bordes*, el proceso estudiado con más amplitud en la visión temprana, se ocupa de esa tarea. Lo hace marcando los bordes que separan los conglomerados de valores de pixel significativamente distintos. El hecho de que los detectores de bordes existan en los niveles inferiores del sistema visual humano es fácil de aceptar, dada la tendencia de nuestros sentidos a preferir cambios sobre estados estables. De hecho, nuestro sistema visual parece diseñado primariamente para detectar cambios en (más que valores

absolutos de) las señales de luz. Imagínese lo que usted vería en una gran superficie blanca y uniforme que ocupa su campo visual: no mucho. La imagen es aburrida porque no hay cambios en la intensidad de la señal a través del espacio. Pero los cambios espaciales en la intensidad no son la única clase de cambios que espera nuestro sistema visual. Si la proyección de la escena compleja que usted está viendo cuando levanta su vista de esta página fuera a quedar completamente quieta en su retina, en pocos instantes quedaría tan privada de rasgos como la mancha blanca. Desaparecería porque no habrían cambios *temporales* en la señal de luz registrada por cada fotorreceptor.

Las células retinianas a las que los fotorreceptores envían sus señales están diseñadas especialmente para detectar cambios en la intensidad de la luz a través del tiempo y el espacio. Las células comparan continuamente los valores de intensidad actuales con los valores de intensidad registrados hace un momento y envían respuestas transitorias que codifican cambios en los valores de intensidad más que los valores de intensidad mismos. Si los valores de intensidad siguen siendo los mismos, la respuesta de las células cae a cero y la imagen se desvanece. Los cambios de intensidad temporales y espaciales están interrelacionados; a medida que el ojo parpadea incansablemente en su cuenca, los bordes en el espacio se mueven a través de los fotorreceptores y se convierten en bordes en el tiempo.

En la visión de máquina, la detección de bordes persigue los cambios de intensidad como los rasgos más importantes y más primarios, y de esta manera ejecuta la primera tarea en la disminución de la masiva cantidad de información que porta la imagen. La dificultad de la detección del borde radica no en encontrar los cambios de intensidad, sino en descartar los cambios que no sean importantes. Los pequeños cambios de intensidad entre conglomerados de pixels son endémicos en toda imagen. Incluso la imagen de una simple hoja de papel blanco no podría transportar el mismo valor para cada pixel a menos que el blanco del papel sea totalmente uniforme, la luz que brillara sobre el papel fuera la misma en todos los puntos y los sensores que registraran la imagen midieran fielmente la cantidad de luz que incide sobre ellos. Estas condiciones no se encuentran en la realidad: la blancura del papel está manchada por impurezas, la luz pega en el papel en algún ángulo, generando sombras, y los

sensores, bombardeados continuamente por fotones al azar, transmiten una señal ruidosa. Un buen detector de bordes debe descontar las subidas y bajadas misceláneas de la intensidad de la señal que surgen de la superficie del papel y detectar solamente los bordes concretos. Idealmente, la detección de bordes debe liberar un dibujo lineal de la escena, capturando los contornos físicos y los límites de los objetos y dejando en blanco las superficies entre ellos. Pero diseñar un detector de bordes para hacer eso no es una tarea tan simple.

Se ha invertido un esfuerzo tremendo para construir detectores de bordes eficientes. El problema para diseñarlos es el movimiento que hay entre que se suaviza el ruido en la señal de intensidad y se definen todos los bordes relevantes. Los buenos detectores de bordes copian la forma en que la retina humana hace esas dos tareas, ejecutándola en dos pasos. Primero, el detector nubla toda la imagen agregando al valor de cada pixel el promedio de los valores de los pixels que están en torno. En segundo lugar, toma la diferencia entre el nuevo valor del pixel y un promedio de los nuevos valores de pixel en un conglomerado circundante mayor, y le asigna el valor de esa diferencia al pixel central.[2] Estas dos operaciones, *nublado* o *filtrado* y *diferenciación*, aseguran que los pequeños cambios en la intensidad de la señal se suavicen (los pequeños cambios se hacen aun más pequeños después del filtrado) y que los cambios grandes resalten. El efecto de esas acciones sobre esta línea que atraviesa la imagen sería que



se transforme en esta otra señal, mucho más clara.



El umbral por debajo del cual se descartan los cambios en la intensidad y por encima del cual se retienen, se puede controlar de diversas maneras: cambiando el tamaño del conglomerado de pixels que se usa para calcular el promedio, cambiando los pesos que los valores de pixel acarrean en promedio y manteniendo sólo los valores borrosos y diferenciados que estén por encima del umbral que se define. Si el umbral es muy bajo, pueden detectarse bordes de intensidad que no representen bordes físicos en la escena, pero si es demasiado alto, algunos bordes reales pueden desaparecer con el ruido. Los diseñadores de detectores de bordes para visión de máquina han puesto un gran esfuerzo en inventar filtros y diferenciadores que destaquen los bordes y suavicen las superficies con efectividad. Ningún detector de bordes individual puede ser perfecto. Sin embargo, los dibujos de líneas que se producen en esta etapa temprana de la visión contienen muchas líneas misceláneas.

Restricciones naturales

El objetivo de la visión temprana de reducir la ambigüedad de las imágenes inspiró la idea de las *restricciones naturales*. Fue obvio desde el comienzo que se necesitaban restricciones para reducir el número de interpretaciones posibles de una imagen, pero lo que no era obvio era de qué ámbito debían tomarse esas restricciones. De allí que se considerara que las *restricciones no naturales* como las del programa de reconocimiento telefónico eran tan buenas como cualquier otra. Al volverse hacia la física de las imágenes los científicos de la visión se volvieron hacia el mundo natural en busca de restricciones: había que establecer supuestos sobre las propiedades de las luces, las superficies y las geometrías que casi siempre fueran verdaderos; en una palabra, había que establecer restricciones naturales. Las restricciones naturales reducen el número de interpretaciones de una imagen excluyendo algunas de ellas como físicamente imposibles.

Un algoritmo estereoscópico. La tarea de encontrar y formular con precisión las restricciones naturales correctas demostró ser la más difícil de la visión temprana. Las restricciones naturales figuran prominentemente en la solución al problema de la percepción de la profundidad.

Usted puede apreciar el problema de la percepción de la profundidad y la forma en que los humanos lo hacen más fácil usando los dos ojos con un simple experimento. Coloque un dedo a varias pulgadas de su rostro y mírelo cerrando el ojo derecho. Ahora abra el ojo derecho y cierre el izquierdo. Su dedo parece saltar hacia la izquierda. Ahora mantenga el dedo tan lejos de su rostro como pueda y alterne la visión entre los dos ojos de la misma manera. Los saltos de su dedo son ahora menores. La explicación: dado que sus ojos ocupan una posición ligeramente distinta, la imagen en cada ojo cae en una forma ligeramente distinta.

La diferencia en la posición de su dedo en las dos imágenes es su disparidad binocular, y, como ha mostrado el experimento, está directamente relacionada con la distancia entre el dedo y el ojo. Cuanto más lejos está el dedo, menor es la disparidad. La visión estereoscópica capitaliza este hecho usando la disparidad para estimar la profundidad relativa de las superficies en una escena y recrear así objetos sólidos y tridimensionales. Para calcular la disparidad binocular de un punto particular de la escena, el sistema visual debe identificar primero qué pixels de las dos imágenes corresponden al mismo punto. Este problema de la correspondencia tiene muchas soluciones: cada pixel en la imagen izquierda corresponde a algunos de los muchos pixels con valores de pixel similares en la imagen derecha. Entremos restricciones naturales, y se eliminarán todas las soluciones excepto las físicamente correctas.

En la búsqueda de restricciones que sean al mismo tiempo poderosas y suficientemente generales para resolver el problema, se mantienen dos supuestos acerca del mundo cotidiano. La restricción del *carácter único* encarna el hecho de que la mayoría de las superficies no son transparentes; esta restricción determina que cada pixel en (digamos) el ojo izquierdo sea asignado a una y sólo una profundidad (esta restricción no se sostendría si la mayoría de las escenas se observase a través de muchas capas de cristal). La restricción de *continuidad* expresa el hecho de que casi todas las superficies son suaves, de manera que si se asigna un punto de una imagen a una cierta distancia desde el ojo, se asignarán a los puntos cercanos distancias similares. Sólo si el mundo estuviera compuesto por chubascos de puntos discontinuos a profundidades al azar, podría esperarse ver profundidades abruptamente distintas para pixels cercanos.

Una vez que se encontraron las restricciones, ¿qué se hace con

ellas? Marr y Poggio[3] las incorporaron a un algoritmo estereoscópico que, en su última versión, comienza con una imagen izquierda y otra derecha, ambas filtradas y diferenciadas por un detector de bordes, y produce una imagen visual en la que cada valor de pixel representa una profundidad. Para cada pixel de la imagen izquierda, el algoritmo computa la disparidad para cada uno de sus posibles correlatos en la imagen derecha. Para cada disparidad de cada pixel, el algoritmo cuenta luego el número de correlaciones para los pixels próximos que exhiben la misma disparidad. La disparidad con el número de «votos» más grande es la que gana. La restricción de continuidad asegura que el método de votar es limpio (si una disparidad es soportada por muchas correlaciones próximas, es probable que sea la correcta en una superficie suave) y la restricción del carácter único determina que haya un solo ganador.

Este método simple de incorporar restricciones naturales consigue resultados realistas con la mayoría de las imágenes. A diferencia de sus predecesores en los días de los *hacks*, el método no descansa en indicios de alto nivel que señalan (digamos) la punta de la nariz en la imagen izquierda, y luego resuelven el problema de la correspondencia reconociendo la punta de la nariz en la imagen derecha. Podemos ver profundidad en estereogramas de puntos al azar, demostrando el hecho notable de que nosotros, al igual que el algoritmo estereoscópico, no necesitamos reconocer los objetos antes de que podamos verlos en profundidad. (Los estereogramas de puntos al azar son imágenes tridimensionales que se crean mostrando al ojo derecho un conjunto de puntos negros al azar sobre un fondo blanco y mostrando al ojo izquierdo otro conjunto. Los patrones de puntos en ambos conjuntos son idénticos, excepto en el hecho de que la mancha central de puntos se halla ligeramente desplazada en la imagen izquierda en relación con su posición en la imagen derecha. El ojo interpreta ese ligero desplazamiento como una diferencia en profundidad de la mancha central y de su trasfondo, de modo que la mancha parece flotar encima o por debajo de los puntos circundantes.)

Optica inversa

Para la investigación en visión temprana, la exploración de las propiedades del ambiente que interactúa con la visión y la búsqueda

de restricciones naturales significa desarrollar una nueva ciencia del mundo: la *óptica inversa*. Así como la óptica es la física de la formación de imágenes bidimensionales a partir de escenas tridimensionales, la óptica inversa es la física de la recuperación de las escenas tridimensionales a partir de las imágenes bidimensionales. Así como los artistas del Renacimiento reproducían contornos y límites tridimensionales sobre un papel bidimensional siguiendo las reglas de la perspectiva lineal, los científicos modernos de la visión están aprendiendo a decodificar imágenes bidimensionales descubriendo las nuevas reglas de la óptica inversa.

La óptica inversa es una ciencia de los problemas imposibles. La información proporcionada en los datos de la imagen es insuficiente, y la solución no puede ser ni única ni bien definida. La IA tradicional puede luchar para definir un problema con precisión, pero una vez que el problema se ha definido, su solución es única y se encuentra en línea recta (aunque consumiendo tiempo). En la óptica inversa lo opuesto es verdad: los problemas son fáciles de estipular pero difíciles de resolver. Uno de los avances más importantes en la visión temprana ha sido la constatación de que esos problemas imposibles encuadran en una clase de problemas (técnicamente llamados problemas mal planteados) que ha sido extensamente estudiada en matemáticas. Dado que los matemáticos ya han desarrollado técnicas útiles para afrontar la forma general de los problemas mal planteados, los científicos de la visión pueden utilizar las mismas técnicas para resolver problemas particulares de óptica inversa. Estas técnicas caen bajo el rubro de la *teoría de la regularización*.[4]

La teoría de la regularización proporciona un marco de referencia en el que las restricciones naturales, una vez descubiertas, se pueden sistematizar. El marco es el mismo para todos los problemas en la visión temprana, y apunta a una solución general que se puede ajustar a cada problema. De este modo, la dificultad de los problemas no-tan-imposibles en la visión radica casi exclusivamente en encontrar las restricciones naturales correctas, ya que ahora se ha prescrito formalmente una forma de ejercitarlas.

Las tácticas de la visión temprana se pueden caracterizar como encuadradas en tres actividades genéricas. La primera es segregar el problema de otros problemas de la visión; es decir, si el objetivo es recuperar el mapa en color de las superficies de la imagen, no hay necesidad de computar la disparidad binocular (nuestras

propias neuronas dividen sus obligaciones de la misma forma: una neurona para seleccionar el color generalmente tiene poco interés por la disparidad). La segunda es identificar las restricciones naturales que gobiernan el problema. La última tarea consiste en encajar las restricciones naturales en un algoritmo que funcione, quizás utilizando el marco de la teoría de la regularización.

Visión de alto nivel

Un «mobot» rueda por un pasillo del laboratorio de IA del MIT, deteniéndose cuando sospecha que hay algo sólido demasiado cerca. Puede seguir una pared o volver atrás sus pasos, buscando su camino como si pudiera ver. Pero su banco de sensores infrarrojos transmite sólo el más básico de los mensajes: hay algo ahí, o no lo hay. No puede decir qué cosa sea.

Aunque mucho más avanzadas que las imágenes del mundo del mobot, el esquema $2^{1/2}$D producido por la visión temprana también se detiene antes de decir qué son las cosas. Su objetivo es determinar dónde están. Determinar qué son las cosas es el objetivo de la visión de alto nivel. Si el esquema $2^{1/2}$D segregaba cada objeto de los demás objetos en la imagen, la tarea de reconocer los objetos por características tales como el color, la forma, la textura, etcétera, no debería ser demasiado difícil. Pero trazar los límites entre objetos separados es exactamente lo que la visión temprana (al menos como la reproducen las máquinas) no puede hacer aún. Las cosas en la imagen que recoge la visión temprana son en el mejor de los casos partes de objetos (superficies no quebradas de un solo color, pequeñas protuberancias oscuras en una superficie lisa) y no los objetos mismos. La dificultad radica no en encontrar los bordes entre regiones de la imagen que son diferentes entre sí, sino en determinar qué regiones son significativas para distinguir y rotular los objetos.

Las tareas de la segmentación de la imagen (calar la imagen en regiones que plausiblemente correspondan a objetos separados) y de reconocimiento de objetos (comparar esas regiones significativas con objetos rotulados en la memoria) han sido el foco de intensos estudios en la visión de alto nivel. Pero aun los programas de reconocimiento de objetos más sofisticados de la actualidad hacen una demanda excesiva a las imágenes que procesan. Los

programas requieren que los objetos que reconocen sean señalados antes en la imagen. Una vez que se le dice así dónde hay que mirar, un programa de esos puede comparar cada rasgo de un objeto señalado con una imagen virtualmente idéntica en su memoria. Pero enfrentado con un esquema $2^{1/2}$D de una escena, el programa no sabe por dónde empezar. El problema que los científicos de la visión enfrentan ahora con avidez es cómo hacer que dos estrategias de la visión (la temprana y la de alto nivel) se encuentren en el medio. ¿Cómo se llega de los bordes a los objetos?

Aunque no se ha contestado aún esta pregunta, la visión de máquina ha hecho un progreso notable desde su nacimiento como ciencia. Cinco años atrás, extraer los bordes de una imagen insumía treinta minutos de tiempo de computadora; ahora toma menos de un décimo de segundo y produce una imagen límpida que los fabricantes de aviones, entre otros, utilizan con provecho. Los algoritmos que recuperan el color, la profundidad, el movimiento y la forma de las superficies de imágenes bidimensionales del mundo funcionan más rápido que nunca. Ya están asistiendo a vehículos militares experimentales autónomos a moverse por la tierra y asisten a los robots industriales en la inspección de productos de fábrica. El próximo obstáculo es integrar distintos algoritmos en un sistema visual que pueda ver en tiempo real. La Máquina de Visión en el laboratorio de IA en el MIT es una primera versión de ese sistema. La visión de máquina también ha tendido puentes a la biología y la psicología y ha demostrado que hacer que las máquinas vean significa mirar dentro de la mente humana.

Niveles de comprensión

El compromiso de la visión de máquina para comprender la visión en todos los niveles surge de la naturaleza y dificultad de los problemas que enfrenta. La visión es un problema duro, y tratar de resolverlo construyendo meros programas de reconocimiento telefónico no funciona. Los logros sólidos de la investigación en visión de máquina se originan en la forma en que se llevó a la práctica su dogma central (que la inteligencia se puede estudiar como un sistema abstracto de procesamiento de la información, independientemente de la maquinaria en que corre), el que se ha convertido en una filosofía y en una ciencia por derecho propio. La

ciencia es la óptica inversa, fundamentada en la física del mundo real y formalizada en términos de una matemática rigurosa. La filosofía que se halla en el corazón de la visión de máquina (especialmente como se la practica en el laboratorio de IA del MIT) está apuntalada por la creencia de que existen niveles de comprensión y de análisis, que dictan que los problemas de procesamiento de la información se resuelvan a tres niveles: computación, algoritmo y hardware. El credo de la visión de máquina es que se ha de encontrar el fundamento más firme para cada problema en el nivel de la computación.

La estrategia computacional sostiene que los problemas de la visión se pueden estudiar como problemas de las matemáticas y de la física, restringidos por las propiedades del mundo al que se refieren las imágenes y del ojo en que se forman. Las soluciones se deben caracterizar completamente con independencia de la maquinaria en que se implementarán; las restricciones naturales que permiten que exista una única solución son las mismas, sea que las exploten neuronas o conmutadores. Como ha escrito David Marr, «una vez que se ha establecido una teoría computacional para un problema en particular, nunca tendrá que ser hecha de nuevo».[5] Esto se ha convertido en uno de los pilares de la IA, igual que como teorema constituye un principio básico de las matemáticas. La teoría computacional que subyace a la detección de bordes, que establece que los rasgos primitivos más significativos en una imagen son los bordes, no está atada a la forma en que cualquier pieza de hardware encuentre los bordes. En vez de eso, es la teoría la que proporciona hechos firmes sobre la información visual.

El algoritmo es el procedimiento paso a paso que ejecuta los comandos de la computación: en la detección de bordes, es el conjunto de instrucciones para calcular la suma de un grupo de valores de pixel, para dividir la suma por el número de pixels en el grupo, para agregar ese guarismo al valor del pixel central, etcétera. El hardware es el conjunto de artefactos que implementa el algoritmo: en el sistema visual humano, las células retinianas están ampliamente interconectadas, lo que permite que las células vecinas alimenten una célula central con la suma de sus actividades. A nivel computacional, la visión de máquina determina lo que necesita computar; a los niveles del algoritmo y del hardware, prescribe cómo realizar la computación.

Cuando Marr y Poggio propulsaron inicialmente el uso del nivel

de computación en la búsqueda de soluciones,[6] enfatizaron su independencia de los otros niveles para separarlo de una caterva de algoritmos de IA. Pero en la práctica los niveles deben interactuar. El algoritmo es dictado por la computación que debe ejecutar y a menudo se halla constreñido por las propiedades y limitaciones del hardware. De modo que posee influencia recíproca, tanto sobre la computación como sobre el hardware. Alterar el procedimiento exacto que sigue un algoritmo (por ejemplo, para aumentar su velocidad o mejorar su confiabilidad) a menudo es lo mismo que modificar la computación que ejecuta. Elaborar un algoritmo puede llevar por ende a la invención de una nueva computación o a la comprensión de lo que es realmente el problema a resolver. De igual manera, el algoritmo puede requerir ciertas manipulaciones (multiplicar mutuamente enormes matrices de números, por ejemplo) que el hardware existente no puede ejecutar con eficiencia, de modo que puede hacer de acicate para la evolución o el descubrimiento de nueva maquinaria. Los esfuerzos en visión de máquina en los últimos quince años demuestran el firme compromiso de atacar la visión en todos los niveles y la creencia en el hecho de que todos los niveles se hallan tan intrínsecamente interrelacionados que se necesita ese compromiso si es que ha de entenderse la visión.

Oposiciones polares

Antes que llegara a ser obvio que la visión era uno de los problemas más duros que podía afrontar la IA, algunos investigadores de IA respondieron a la exhortación de Marr de construir sólidas teorías computacionales antes que algoritmos desvencijados con la argumentación mordaz de que eso estaba bien para cosas sencillas, como la visión, pero no para las cosas más difíciles, como la alta inteligencia. Hoy, el impacto que la filosofía de la visión de máquina ejerce sobre la IA tradicional se halla reforzado por el conflicto de la IA con el conexionismo. Aunque encontramos que el conexionismo también puede aprender algo de la visión de máquina.

La IA y el conexionismo son dos ramas de la misma empresa; se las puede ver como si representaran los dos polos de la inteligencia. La filosofía conexionista se inspira en nuestros

poderes *asociativos* (asombrada por la forma en que nosotros vadeamos a través de la arena movediza de las restricciones múltiples, hablando, canturreando, manejando vehículos, buscando tazas de café, reconociendo rostros en la multitud), mientras que la IA se inspira en nuestros poderes *deductivos* (impresionada por la lógica, las pruebas matemáticas, los debates legales y la eliminación sistemática de posibles errores en el código de las computadoras). Sus diferentes visiones de la inteligencia conducen a planes diferentes para recrearla:

• La IA se alucina con los algoritmos; los conexionistas insisten en el hardware, El conexionismo sostiene que los algoritmos por sí solos no pueden recrear la inteligencia y que el énfasis de la IA en los algoritmos otorga una primacía inapropiada al procesamiento simbólico, el cual nunca puede capturar la «fluidez y adaptabilidad»[7] de la inteligencia humana.

• El hardware es la esencia de la inteligencia, dice el conexionismo, y no sólo la IA ha olvidado este hecho, sino que utiliza el hardware inadecuado. La IA ha prosperado por el rápido desarrollo de computadoras seriales más poderosas, «máquinas de von Neumann», que desarrollan instrucciones una después de la otra. Los conexionistas creen que el hardware debe ejecutar las operaciones no en serie, sino en paralelo, y que las cantidades con las que debe trabajar son números, y no símbolos, y analógicos, y no digitales. (En una computadora digital los datos están representados y son operados como hileras de ceros y unos —dígitos binarios—. En una computadora analógica los datos están representados como cantidades físicas, tales como voltajes, que asumen rangos continuos de valores.)

• Los conexionistas creen además que el hardware correcto es una red densamente conectada de unidades simples que procesan simultáneamente partes interactuantes del mismo problema. La salida de ese sistema está determinada por la suma de la actividad total de la actividad de todas las unidades de la red, y no por los valores «sí» o «no» de un predicado singular que termina una serie de deducciones lógicas.

• El sueño final de los conexionistas y de los investigadores en IA es construir una máquina que pueda aprender. Los conexionis-

tas predicen que el hardware correcto se organizará a sí mismo espontáneamente (quizá mágicamente) en un sistema que sea inteligente, no simplemente en virtud de lo que se le dijo que hiciera, sino porque puede aprender y generalizar a partir de ejemplos. Contendrá los elementos mismos de la mente que en agregado (así como las moléculas de agua coalescen en copos de nieve) exhibirán propiedades emergentes tales como la inteligencia. Alimenten la clase adecuada de red con una lista de palabras escritas y sus pronunciaciones correctas, y esa red se dará cuenta del estado en que debe estar para pronunciar palabras que no estén en su lista de entrenamiento. La computación que la red ejecuta para ir del texto al habla no necesita explorarse.

Más que a cuestiones de hardware contra software, de símbolos contra números o de operaciones seriales contra operaciones paralelas, el debate decanta en una sola cuestión: ¿cuál es el objetivo final de la empresa? O, dicho de otra manera, ¿cuál es el objetivo de estudiar la inteligencia?, ¿construir máquinas inteligentes?, ¿comprender cómo está construido el cerebro? ¿o describir la estructura y los poderes de la inteligencia como una entidad en flotación libre, no ligada ni al cerebro ni a la máquina?

Si inventamos una conexionista ideal y un hacker ideal de IA y le hacemos a cada uno esta pregunta, obtendremos respuestas tajantemente discrepantes. La conexionista ideal responde que su objetivo es construir un modelo del cerebro mediante la simulación de sus redes neuronales. El modelo capturará lo suficiente del poder natural del cerebro como para ser comercialmente viable. Ella se apartaría de una teoría de la inteligencia en formación libre para evitar las «intermediaciones» de símbolos que se filtran en la transacción entre los datos y la solución. En la práctica, las unidades en la mayoría de las redes conexionistas se han simplificado al punto en que no se parecen en nada a las neuronas reales, que son biofísica y computacionalmente dispositivos muy complejos. Un conexionista real admitiría que el único parecido verdadero entre las redes neuronales artificiales y el cerebro se halla en el nivel abstracto de los miles de conexiones y los miles de operaciones simultáneas.

El hacker ideal de la IA, por otra parte, puede afirmar que es el primero en querer construir una máquina inteligente, pero probablemente vacilaría en decidir si es esencial sondear en un cerebro húmedo. Eso no es decir que el hacker rechazaría las

intuiciones ofrecidas por los científicos que sondean cerebros. Pero el hacker dice que él toma en serio el objetivo excelso de comprender la inteligencia puramente como un sistema abstracto de procesamiento de la información, como prueba de que él produce sólo programas de computadora específicos para cada tarea.

En términos de niveles de comprensión, la IA profesa estar en el nivel computacional, mientras que en realidad está pegada al nivel algorítmico.[8] El conexionismo profesa ignorar el nivel computacional y que sólo intenta construir hardware como el del cerebro. Pero el hardware de las redes del conexionista está muy lejos del cerebro, y muchas de las redes trabajan sólo porque el análisis computacional necesario se ha hecho primero. (Por ejemplo, la red de John Hopfield es simplemente una máquina de minimización. Es decir, antes de usarla para resolver un problema, uno debe expresar el problema, si es posible, como una cantidad matemática que debe minimizarse. Este análisis preliminar se encuentra en el nivel computacional y tiene poco que ver con la red misma.) El mensaje de la visión de red, tanto al conexionismo como a la IA, es que ninguno de sus objetivos podrá alcanzarse sin la persecución simultánea del otro.

Visión: una síntesis

En realidad, los límites entre la IA tradicional y el conexionismo no están tan gruesamente marcados. Aunque sus doctrinas y técnicas parezcan tan diametrales como los polos de la inteligencia que los inspiran, ambos convergen en la visión de máquina. Mientras la conexionista ideal y el hacker de IA ideal se oponen, el científico de la visión de máquina sigue un curso constante que coincide con partes de sus respectivas estrategias.

La facilidad, inmediatez e inescrutabilidad de la visión se ubica en el polo asociativo de la inteligencia. Con la gruesa resolución de nuestra conciencia, así es como lo vemos. Pero en la escala fina entrevista por las teorías computacionales, la visión a menudo opera de una manera sumamente deductiva. A su vez, la visión de máquina adopta metodologías contrastantes inspiradas por los dos polos de la inteligencia: desarrolla algoritmos altamente paralelos, trata afanosamente con números crudos, descansa en

teorías abstractas de procesamiento de la información y ensambla sistemas expertos visuales.

La visión de máquina es el juego más numérico y paralelo que se juega en la ciudad. El objetivo de la mayoría de los procedimientos en visión de máquina es transformar una enorme matriz de números en otra matriz también enorme de números, y no evaluar la verdad lógica de una sola proposición. Dada la enormidad de la primera matriz, y dado que todos los puntos en ella a menudo deben ser transformados de la misma manera (por ejemplo, un detector de bordes ejecuta exactamente la misma operación, sin que importe cuál es el conglomerado de pixels de que se trate), la forma más natural de hacer esta transformación es simultáneamente, en paralelo, sobre cada pixel. La mayoría de los algoritmos iniciales de la visión de máquina se han desarrollado con la idea del procesamiento paralelo en mente, utilizando la retina y el cerebro (órganos prototípicamente «paralelos») como modelos.

El primer artículo de Marr y Poggio sobre la visión estereoscópica (1976)[9] comienza: «Quizás una de las diferencias más chocantes entre un cerebro y las computadoras actuales es la cantidad de 'cableado'. En una computadora digital la relación entre las conexiones y los componentes es del orden de 3, mientras que en la corteza de los mamíferos está entre los 10 y los 10.000. Aunque este hecho señala una clara diferencia estructural entre ambos, la distinción no es fundamental para la naturaleza del procesamiento de información que cada uno realiza, sino meramente respecto de los pormenores de la forma en que cada uno lo hace. En términos de Chomsky, esta diferencia afecta a las teorías de la performance pero no a las teorías de la competencia, porque la naturaleza de una computación realizada por una máquina o un sistema nervioso depende sólo del problema a resolver, y no del hardware disponible. Sin embargo, puede esperarse que un sistema nervioso y una computadora utilicen diferentes clases de algoritmos, incluso cuando ejecutan la misma computación subyacente. Los algoritmos con una estructura en paralelo, que requieren muchas operaciones simultáneas locales en grandes matrices de datos, son caros para las computadoras actuales, pero probablemente adecuados para la organización altamente interactiva del sistema nervioso...»

Muchos de esos algoritmos, aunque probados inicialmente en computadoras digitales, se pueden implementar fácil y más

eficientemente en forma altamente paralela en redes cuyas actividades por unidad se expresen en cantidades analógicas; la teoría de la regularización, que unifica muchos de los algoritmos de la visión temprana, muestra una forma natural de hacerlo.[10] La Máquina de Conexión, una poderosa computadora que consiste en muchos miles de procesadores simples densamente conectados, fue originariamente concebida, en parte, para la investigación de la visión.

La visión de máquina también ha permanecido cerca del cerebro. Los investigadores de la visión se han dado cuenta de que las máquinas artificiales tienen mucho que ganar emulando las máquinas naturales, que se comportan tan extraordinariamente bien en la visión y en los demás sentidos. De acuerdo con eso, han modelizado detectores de bordes según las células retinianas y han tomado otros indicios del cerebro. Mientras tanto, los biólogos del cerebro miran cada vez más hacia la visión de máquina en busca de intuiciones en torno de las operaciones que deben realizar las neuronas para resolver los problemas perceptuales que enfrentan.

Pero la visión de máquina no implora exactamente a la IA que se vuelva hacia los números y el paralelismo. Después de todo, la IA no puede simple e instantáneamente transformar las bases de datos y las máquinas de inferencia de los sistemas expertos en enormes matrices de números.[11] Ni la visión de máquina instruye al conexionismo para que sea más fiel al cerebro. Lo que es más importante, pide tanto al conexionismo como a la IA tradicional buscar soluciones a nivel computacional.

Una solución conexionista al problema de la visión estereoscópica puede consistir en alimentar conjuntos de tres imágenes en una red artificial de neuronas, una imagen de entrada para cada ojo (lo que da la intensidad de luz para cada pixel) y una imagen de salida, la solución (que da la distancia del observador para cada pixel). Para cada conjunto, la red produciría sus propias imágenes de salida a partir de los dos insumos, las compararía con la solución correcta y ajustaría las fuerzas de las conexiones entre sus unidades para hacer que las dos imágenes de salida coincidieran. Con suficientes conjuntos de entrenamiento, la red podría con el tiempo definir un patrón de fuerzas que produciría una imagen de profundidad adecuada cuando se la alimenta con un par completamente nuevo de imágenes de entrada. Mirar dentro

de la red puede revelar una masa de conexiones excitatorias e inhibitorias muy parecida a la que el algoritmo de Marr y Poggio define para resolver el problema. Pero mientras el algoritmo de Marr y Poggio surge de un análisis computacional de la percepción de la profundidad, y por lo tanto funciona dentro de un dominio completamente descrito, sólo se puede conjeturar qué y cuánto es lo que podrá hacer la red conexionista.

¿Cómo podría la IA beneficiarse de la estrategia computacional? Tomemos los hábitos de vuelo de la mosca doméstica ordinaria. Los gustos de la mosca para determinar sus blancos son simples porque su visión es tosca. Cualquier patrón blanco y negro (una miga en un mantel) o un punto negro que se mueve (otra mosca, posiblemente del sexo opuesto) puede alertar a la mosca para que la siga. Un hacker tradicional de IA estaría tentado a construir reglas explícitas que dicen a la mosca cómo seguir a una pareja potencial: *Si el punto negro gira a la derecha, doblar a la derecha. Si el punto negro revolotea, interceptarlo.* El esquema detrás de las reglas sería comparar continuamente la dirección del blanco y la dirección de la mosca, y hacer movimientos para emparejarlos. La estrategia computacional podría desarrollar un esquema similar, pero no empaquetando un conjunto de reglas para cubrir sólo una lista finita de maniobras.

La teoría de Poggio-Reichardt[12] sobre el vuelo de las moscas es una teoría computacional clásica. La teoría pone aparte las reglas, considerando sus estructuras subyacentes. Especifica las reglas en un solo enunciado matemático que muestra la forma en que los insumos visuales se transforman en salidas motoras, restringidas por la física del vuelo y la biología de la mirada. Los enunciados ecualizan el impulso de rotación ejercido por las alas de la mosca (que a su vez controla su posición y velocidad) con la diferencia entre la posición actual y la posición deseada de la imagen del objetivo en la retina. La rapidez del cambio de la diferencia determina la velocidad del cambio del impulso de rotación. Una sola ecuación sintetiza el patrón del vuelo de la mosca en una persecución.

Aunque un conexionista puede encontrar que la conducta de la mosca es groseramente poco sofisticada, puede demostrarse que el acto humano reflejo de pisar los pedales cuando un vehículo resbala frente a una señal de alto está gobernado por la misma ecuación. Pero el mayor impedimento en el programa de la mosca

de los hackers es más elocuente que la denigración conexionista de la mosca: el programa de IA tradicional presupone que el trabajo duro de seleccionar el objetivo ya ha sido hecho. Sus reglas sólo se aplican cuando se suministra información de alto nivel sobre la ubicación y velocidad de las manchas negras. Si fuera un programa verdaderamente inteligente, comenzaría con la imagen sin elaborar producida por el ojo primitivo de la mosca, encontraría en él los puntos prominentes y seguiría esos puntos continuamente en el tiempo. El trabajo radica en descifrar los errores en la posición retiniana del blanco y en descubrir que el impulso de rotación del ala es la salida relevante. De igual modo, la red conexionista de visión estereoscópica perdería mucho tiempo hasta llegar a un estado productivo si sólo se le suministraran las imágenes sin elaborar. La masa de información en esas imágenes crudas sería demasiada para una red artificial a menos que la red misma fuera imprácticamente grande y compleja. La red de visión estereoscópica lo haría mucho mejor con las imágenes con bordes detectados que utiliza el algoritmo de Marr y Poggio para recortar el número de equivalencias posibles entre los pixels de ambas imágenes. Para trabajar correctamente, el programa de la mosca y la red de visión estereoscópica requieren ambos la representación correcta de los datos de entrada.

Lo que hace una teoría computacional es encontrar la representación correcta. Esta le dice cómo ir del insumo a la salida encontrando el insumo y la salida adecuados. Especifica tanto la información como la forma de procesarla. En la búsqueda de los elementos básicos de la cognición, la IA ha buscado los pasos mínimos de deducción en las reglas más globales del pensamiento, mientras que el conexionismo ha buscado los nexos más recónditos de la asociación. La visión de máquina ha vuelto su mirada hacia afuera. Ella busca generalizaciones sobre el mundo que sean casi siempre verdaderas (los objetos son rígidos, las superficies son suaves, los límites son continuos) y las traduce en restricciones sobre los elementos básicos de la información.

La visión de máquina comparte el sueño de construir una máquina que pueda aprender. Pero hay algunas preguntas que deben contestarse primero. ¿Es posible aprender alguna computación a partir de un conjunto de ejemplos, comenzando con una tabula rasa? Pensamos que probablemente no. Para la mayoría de los problemas, el marco de referencia que guía los datos hacia una

solución debe existir antes que el aprendizaje pueda actuar sobre él, acelerando y perfeccionando la solución. Ciertas transformaciones de los datos en soluciones probablemente no puedan aprenderse en absoluto, excepto mediante una búsqueda exhaustiva de todas las soluciones posibles. Pero a pesar de su respuesta, la pregunta representa una veta en la mina de la investigación. Lo que debe extraerse de la excavación son caracterizaciones de las computaciones que hay que aprender, y cuán bien y mediante qué clases generales de redes pueden ser aprendidas.[13] En la actualidad se han hecho muy pocas excavaciones. Pero algunos científicos de la visión de máquina están explorando las implicaciones de la teoría de la regularización, la que muestra que bajo ciertas circunstancias, algunos algoritmos de la visión se pueden aprender a partir de ejemplos.

¿Quedaremos satisfechos simplemente construyendo máquinas que puedan aprender? Desde un punto de vista práctico, la respuesta probablemente sería que sí; las máquinas serían por cierto muy útiles. Pero si el objetivo es comprender la inteligencia, la respuesta es no. Reproducir simplemente una habilidad no explica sus estrategias subyacentes. Los humanos pueden aprender, pero no saben cómo. La teoría de la evolución proporciona una descripción completa y consistente de la forma en que se desarrolló la vida (y los cerebros). Ella nos dice cómo construir un sistema nervioso, aunque por desdicha el procedimiento insume demasiado tiempo como para ser práctico. Pero esta teoría de la vida y la inteligencia, como la perspectiva que dice que es suficiente construir una máquina inteligente, no es suficiente para los que quieren comprender qué es la inteligencia. De la misma manera, aun si se descubre que la red mágica puede aprender a resolver cualquier problema, un verdadero creyente en los niveles de comprensión insistiría en seguir preguntando: ¿cómo es que la red ha aprendido?

No se debe acusar a los humanos de que son usados como prueba de la existencia de máquinas procesadoras de información, sólo porque han rechazado la manipulación simbólica de los programas tradicionales de IA. Aunque ese estilo de programación sigue siendo más adecuado a las preguntas que ha atacado tradicionalmente en materia de razonamiento, resolución de problemas y lógica, no puede mantenerse por sí mismo. Así como el pensamiento tiene dos sabores y la inteligencia dos polos, el

estudio de la inteligencia debe abrevar en dos filosofías. Actos tan profundamente inteligentes como la percepción, el reconocimiento del habla y el control motor necesitan una estrategia más numérica, paralela y analógica. Nosotros los humanos no debemos olvidar que los que buscan construir máquinas inteligentes tienen todo el futuro por delante para desaprobar su hipótesis de partida: que la inteligencia se puede reproducir en una máquina. Hoy la inteligencia humana excede con mucho las capacidades de los sistemas expertos o de las redes conexionistas, pero en el futuro, máquinas más sofisticadas pueden ofenderse ante esa afirmación. Esas máquinas podrían mirar hacia los días en que la visión de máquina, combinando todos los niveles de comprensión de la inteligencia humana, puso a sus antecesores lado a lado.

Notas

[1] Para definiciones más extensivas de símbolos, estructuras simbólicas, designación e interpretación véase Allen Newell y Herbert A. Simon, «Computer Science as Empirical Inquiry», en *Mind Design*, edición de John Haugeland (Cambridge: Bradford Books, MIT Press, 1981).

[2] En realidad, la forma en que se determina la diferencia es un poco más complicada, pero el efecto es similar. Para una descripción más detallada de la detección de bordes, véase Berthold K. P. Horn, *Robot Vision* (Cambridge: MIT Press; Nueva York: McGraw-Hill Inc., 1986).

[3] David Marr y Tomaso Poggio, «Cooperative Computation of Stereo Disparity», *Science* 194 (1976):283-87.

[4] Tomaso Poggio, Vincent Torre y Christof Koch, «Computational Vision and Regularization Theory», *Nature* (1985):314-19.

[5] David Marr, «Artificial Intelligence: A Personal View», en *Mind Design*, ed. por Haugeland.

[6] David Marr y Tomaso Poggio, «From Understanding Computation to Understanding Neural Circuitry», en *Neuronal Mechanisms in Visual Perception*, edición de E. Poppel, R. Held y J. E. Dowling, *Neuroscience Research Progress Bulletin* 15 (1977): 470-88.

[7] David E. Rumelhart, James L. McClelland y el PDP Research Group, *Foundations*, vol. 1 de *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Cambridge: MIT Press, 1986), 3.

[8] Daniel Demmet hace una aseveración similar en su artículo «The Logical Geography of Computational Approaches: A View from the East Pole», presentado en la Conferencia de Filosofía y Ciencia Cognitiva en el MIT, 17 al 20 de mayo de 1984.

[9] Marr y Poggio, «Computation of Stereo Disparity».

[10] Berthold Horn fue probablemente el primero (en 1974) en usar redes analógicas para resolver un problema de visión, la computación de la luminosidad (véase Horn, *Robot Vision*). Para la conexión entre las redes analógicas y los algoritmos de la visión temprana véase Poggio et al., «Computational Vision and Regularization Theory».

[11] Algunos conexionistas están tratando de hacer exactamente eso; vean si no las bases de datos médicas creadas por James Anderson y codificadas en redes.

[12] Werner Reichardt y Tomaso Poggio, «Visual Control of Orientation Behaviour in the Fly», *Quarterly Review of Biophysics* 9 (1976):311-438.

[13] Hay diversas preguntas básicas que surgen de la estrategia conexionista frente al aprendizaje, pero no han sido contestadas por ella. Las técnicas conexionistas de aprendizaje, tipificadas por el ejemplo de la visión estereoscópica, ¿sólo funcionan para problemas de pequeño tamaño? ¿Escalan adecuadamente a problemas de mayor tamaño? Más fundamentalmente, ¿qué tipos de aprendizaje tienen probabilidades de funcionar en qué clases de problemas? Finalmente ¿son los algoritmos conexionistas de aprendizaje significativamente distintos de las técnicas clásicas de regresión y conglomerado? Aventuramos la respuesta de que pueden no serlo.

Lecturas complementarias

Ballard, Dana H. y Christopher Brown. *Computer Vision.* Englewood Cliffs, Nueva Jersey: Prentice-Hall, Inc., 1982.

Barrow, Harry G. y Jay M. Tenenbaum. «Computational Vision». *Proceedings of the IEEE* 69 (5) (1981).

Charniak, Eugene y Drew McDermott. *Introduction to Artificial Intelligence.* Reading, Mass.: Addison-Wesley Publishing Co., 1985.

Grimson, W. Eric L. *From Images to Surfaces: A Computational Study of the Human Early Visual System.* Cambridge: MIT Press, 1981.

Horn, Berthold K. P. *Robot Vision*, Cambridge y Nueva York: MIT Press y McGraw-Hill, Inc., 1986.

Marr, David. *Vision.* San Francisco: Freeman, Cooper & Co., 1982.

Rumelhart, David E., James L. McClelland y el PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Cambridge, MIT Press, 1986.

Ullmann, Shimon. *The Interpretation of Visual Motion.* Cambridge: MIT Press, 1979.

Winston, Patrick. *Artificial Intelligence.* Reading, Mass.: Addison-Wesley Publishing Co., 1984.

11
Inteligencia artificial y psicoanálisis: una nueva alianza

Sherry Turkle

La inteligencia artificial y el psicoanálisis parecen ser mundos aparte. El psicoanálisis persigue lo que es más humano: el cuerpo, la sexualidad, lo que se sigue del hecho de haber nacido y haber sido criado en una familia. La inteligencia artificial busca deliberadamente lo que es menos específicamente humano: la fundamentación de su visión teorética es la tesis de que la esencia de la vida mental es un conjunto de principios que pueden compartir la gente y las máquinas.[1]

Hay otro aspecto en el que parecen mundos aparte. La inteligencia artificial parece científicamente en ascenso y ha determinado cada vez más la agenda de la psicología académica a través de su influencia sobre la psicología cognitiva. En contraste, el psicoanálisis es rechazado por la psicología académica y se halla en conflicto con las tendencias biológicas dominantes en psiquiatría. Aunque ha habido recientemente rachas de interés en la teoría freudiana, ellas han venido del mundo del análisis literario y la filosofía. Para los círculos científicos, el psicoanálisis parece una disciplina congelada: congelada en el lenguaje científico de otra época, congelada en los supuestos psicológicos de otra cultura.

En este ensayo propongo que si el psicoanálisis se halla en problemas, la inteligencia artificial puede ser capaz de ayudarlo. Y sugiero la naturaleza de esta ayuda argumentando que una de las formas en que las computadoras influyen en el pensamiento psicológico es a través de una ruta que no es esencialmente

---

Sherry Turkle. Profesora asociada de sociología en el programa en Ciencia, Tecnología y Sociedad del Instituto Tecnológico de Massachusetts. Es autora de Psychoanalytic Politics: Freud's French Revolution *y de* The Second Self: Computers and the Human Spirit.

técnica. Más bien, las computadoras proporcionan a las ciencias de la mente una clase de legitimación teorética que yo llamo mitos sustentatorios. Por cierto, el impacto inicial de la computadora en la psicología fue claramente de esta naturaleza.

Mitos sustentatorios

Tan recientemente como en la década de 1950, el conductismo dominaba la psicología académica norteamericana; su espíritu se hallaba capturado por la afirmación de que era permisible estudiar la memorización, pero se consideraba una violación del rigor científico hablar de «la memoria». Se podía estudiar la conducta, pero no los estados interiores. En la jerga actual, lo que había en el medio era una caja negra que no debía abrirse ni siquiera especulativamente.

Hacia fines de la década de 1960, la hegemonía conductista se había quebrado, lo mismo que las inhibiciones sobre el estudio de la memoria y de los procesos interiores de la mente. Ciertamente, dentro de la psicología académica apenas subsistían huellas de la metodología conductista. El conductismo no había sido refutado por un experimento crítico. Había muchos factores que incidían en esta revolución científica, incluyendo el clima político y cultural de los años 60. Y el más importante fue la computadora.

El papel de la computadora en la caída en desgracia del conductismo no fue técnico. Fue la misma *existencia* de la computadora lo que proporcionó la legitimación a una forma radicalmente distinta de ver la mente. Los computadores científicos habían necesitado desarrollar un vocabulario para hablar sobre lo que sucedía dentro de sus máquinas, los «estados internos» de los sistemas generales. Si las «mentes» de las nuevas máquinas tenían estados internos, seguramente la gente también. El psicólogo George Miller, que estaba en Harvard durante el apogeo del conductismo, ha descrito la forma en que los psicólogos comenzaban a sentirse embarazados por no serles permitido discutir sobre la memoria ahora que las computadoras tenían una:

Los ingenieros nos mostraron cómo construir una máquina que tenía memoria, una máquina que tenía un propósito, una máquina que jugaba ajedrez, una máquina que podía detectar señales en presencia de ruido,

etcétera. Si ellos podían hacerlo, entonces la clase de cosas que podían decir sobre las máquinas deberían serle permitidas a los psicólogos decirlas de los seres humanos.[2]

La presencia de la computadora volvió a legitimar el estudio de la memoria y de los estados interiores en la psicología científica. Muchos conceptos técnicos que los psicólogos tomaron de la computación (ideas de la cibernética y de la teoría de autómatas) han existido antes que las computadoras se hicieran más seductoras por causa de ellas. «De repente», dice Miller, «los ingenieros estaban utilizando los términos mentalistas que los psicólogos de corazón tierno deseaban usar, pero de los que les habían dicho que eran anticientíficos.»[3] Las ideas computacionales, el lenguaje computacional y la presencia física de las máquinas crearon un clima intelectual en el que era permisible hablar de los procesos mentales proscritos por el conductismo. La presencia de la computadora sirvió de mito sustentatorio para una nueva psicología de los estados interiores que llegó a conocerse como la ciencia cognitiva.

Los programas de computadora proporcionaron una vía para discutir las creencias y las reglas como causantes de la conducta. ¿Por qué el peón comió al peón? Los psicólogos de la vieja guardia hubieran rechazado «porque el peón bloqueó al alfil» como una explicación causal. Eso sería meramente exponer las «razones» del jugador de ajedrez. Pero si la mente es un programa, las razones se transforman en explicaciones. Gran parte del atractivo que la computadora ejerce sobre los psicólogos radica en que ella les permite abrir esa caja negra que es la mente. Una vez que la caja se abre, la computadora sugiere formas de llenarla con conceptos próximos a los de la comprensión de sentido común.

Por cierto, lo esencial de la historia de Miller sobre la memoria es que las computadoras dieron a los psicólogos permiso para investigar algo que «todo el mundo sabe», pero que había sido proscrito de la ciencia: la idea de que la gente tiene memorias, recuerdos. En las últimas dos décadas, la ciencia cognitiva ha estado legitimada por la computadora, legitimando el estudio de algo más que «todo el mundo sabe», esta vez la idea de que la gente posee información y utiliza reglas y que gran parte de esa información se puede formular en palabras. A fines de la década de 1950 Allen Newell y Herbert Simon construyeron un programa de computadora llamado el General Problem Solver (GPS), el cual

estaba orientado por algo que se parecía mucho a las razones, recodificadas como reglas computacionales. Preguntas tales como «¿Por qué el GPS hace tal y tal cosa?» podían responderse con referencia a las reglas que se le habían dado. ¿Por qué no se habrían de usar referencias a reglas para contestar preguntas sobre lo que la gente hace cuando enfrenta problemas similares? La existencia del GPS otorgaba credibilidad a esta pregunta.

Existe una concepción muy difundida en el sentido de que la presencia de la computadora tiende a empujar a la psicología hacia teorías más rigurosas y cuantificables, argumentándose que la computadora, por su naturaleza, requiere reglas, rigor y formalismo. Pero la historia de la influencia de la computadora en la psicología no es tan simple. Por ejemplo, su «primer acto» (el ataque contra el conductismo) fue en dirección a la creación de una ciencia de la mente menos y no más rigurosa, una ciencia más «blanda» antes que más «dura».

La inteligencia artificial es el canal más explícito para la influencia de la computadora sobre la psicología. Ella reafirma un materialismo global y ofrece además teorías particulares sobre la forma de funcionamiento de la mente. Su agenda dual es construir «máquinas que piensen» y usar las máquinas para pensar acerca del pensamiento. Su premisa metodológica es que si uno construye una máquina que pueda hacer algo inteligente, la forma en que se obtiene que la máquina lo haga es relevante también para pensar la forma en que la gente lo hace.

Habitualmente se considera que la inteligencia artificial tiene sus afinidades más firmes con las filosofías racionalistas, que definen el conocimiento como información y que devalúan la ambigüedad mientras ofrecen una concepción de la mente como programa. Pero esta perspectiva, un lugar común en la cultura literaria, es sólo parcial. Se la puede llamar el estereotipo literario del campo. Pero la IA posee otras dimensiones que le dan un rango mucho más amplio de conexiones intelectuales y de implicaciones para la psicología.

Este ensayo es interpretativo e intenta identificar tendencias e influencias concretas de la computadora sobre el pensamiento psicológico; y es especulativo, prediciendo una nueva alianza entre la IA y el psicoanálisis, un compañero intelectual harto alejado del racionalismo, la cuantificación y las proposiciones formales. Esto no significa que haya una identidad de espíritu entre los modelos computacionales y psicoanalíticos, o incluso que no haya incompa-

tibilidades fundamentales entre ellos. Pero aparte de una lista de afinidades tradicionales entre el psicoanálisis y la IA, ellos tienen en común algo nuevo. En años recientes, los computadores científicos y los psicoanalistas han hablado un lenguaje sorprendentemente similar sobre los agentes internos que construyen el pensamiento y el sujeto sensitivo. Detrás del lenguaje hay preocupaciones compartidas que sugieren nuevos nexos teóricos y una nueva fuente de vitalidad para las ideas psicoanalíticas.

Las predicciones siempre son peligrosas. La pertinencia de formular una aquí tiene que ver con lo que se gana al sostenerla: comprensión de la forma en que la presencia de la computadora puede actuar como un mito sustentatorio para fundamentar no ya una cultura psicológica, sino una variedad de ellas.

Afinidades tradicionales

La idea misma de la IA (crear mente en las máquinas) subvierte las nociones tradicionales del sujeto autónomo de una manera que se asemeja a la empresa psicoanalítica. La mayoría de la gente concibe al sujeto autónomo como una idea no problemática porque posee la experiencia cotidiana de tener uno. Nuestro lenguaje cotidiano captura la experiencia y expresa la idea del libre albedrío; decimos «yo actúo», «yo hago», «yo deseo». E incluso cuando la gente ha aprendido a través de la teología o la filosofía a cuestionar la idea del libre albedrío, lo que tiende a hacer es a introducir pequeñas modificaciones en su noción del sujeto autónomo; éste se convierte en un sujeto cuyas decisiones se encuentran restringidas. Inherente al psicoanálisis, es una duda más radical. El inconsciente no restringe; constituye un sujeto descentrado. Inherente a la IA, es un desafío aun más amenazador: si la mente es un programa, ¿dónde está el sujeto? Esto pone en cuestión no solamente si el sujeto es libre, sino si hay al fin de cuentas un sujeto.

El humanismo tradicional se halla comprometido con la idea de que hay un sujeto actuante intencional. En su desafío al sujeto humanístico, la IA es subversiva de una manera que la aparta del racionalismo y la pone en compañía del psicoanálisis y de escuelas filosóficas tales como el desconstruccionismo. El sujeto psicoanalítico está descentrado en la hebra del inconsciente; el sujeto desconstruccionista está descentrado en el lenguaje; el sujeto computacio-

nal está descentrado (por cierto, quizá disuelto) en la idea de programa.

Estas afinidades no reaseguran al humanista tradicional que se ha acostumbrado a ver la IA como un enemigo. Esas afinidades no hacen que la IA sea menos que un ataque contra la idea del sujeto. Pero este ataque proviene de la izquierda, por así decirlo, más que de la derecha. La inteligencia artificial ha de ser temida como lo son Freud y Derrida, y no como lo son Skinner y Carnap.

La «explicación» computacional del ajedrez otorga puntos a otra forma en la que la IA es más como Freud que como Skinner. En la ciencia tradicional, y ciertamente para el conductismo, la línea entre el sujeto y el objeto se considera como sagrada. Pero para Freud su autoanálisis, su técnica de autocomprensión, era indisociable del desarrollo de su teoría general. Como el psicoanálisis, los teóricos de la IA han hecho una profesión de disolver la línea entre la reflexión objetiva y subjetiva. La inteligencia encarnada en el ajedrez es inteligencia que se deriva del conocimiento personal del ajedrez. «Hay sólo un lugar en el cual obtener ideas sobre la inteligencia, y es en el pensamiento sobre mí mismo», dice el científico de la IA Roger Schank. «En última instancia, yo me tengo a mí mismo, y si eso se siente bien, eso es en lo que debo confiar», dice Donald Norman, un psicólogo cognitivo influenciado por la IA.[4]

Marvin Minsky, uno de los fundadores y líderes teóricos de la IA durante el último cuarto de siglo, siempre ha puesto en claro que, por lo que él sabe, sólo se puede hacer que una máquina haga lo que uno mismo sabe cómo hacer. Para construir un programa, usted tiene que sumergirse en una actividad autoanalítica. A comienzos de la década de 1960, Minsky trabajó con un estudiante, Thomas Evans, en un programa de IA que pudiera pasar los tests familiares de analogía visual: *A* es a *B* como *C* es a *D*, *E* o *F*, donde cada letra está en lugar de un dibujo geométrico. Su método era psicológico: piensa acerca de ti mismo. Y su punto de referencia era el psicoanálisis: «Lo que tenían que hacer era algo parecido a lo que hizo Freud. Tom Evans y yo nos preguntábamos a nosotros mismos, en profundidad, qué haríamos para resolver problemas como éste, y eso parecía funcionar bastante bien».[5]

El conductismo prohíbe rigurosamente cualquier referencia a la experiencia personal, y la mayoría de las demás escuelas psicológicas procura ignorar el asunto. Pero la IA y el psicoanálisis han articulado la necesidad de integrar la referencia personal en

la construcción teórica. Cada una, a su propia manera, es una ciencia de la autorreflexión.

Pero ¿son superficiales esas afinidades? Después de todo, el psicoanálisis explora la mente para descubrir lo irracional; la inteligencia artificial inventa máquinas mediante la explotación de lo racional. De hecho, lo que hay entre el psicoanálisis y la IA no es el «materialismo» de la IA. En el pasado cuarto de siglo, el psicoanálisis ha aprendido la necesidad y la productividad de un diálogo intensificado con la farmacología y la neurociencia. Y Freud mismo esperaba que su ciencia de la mente se aferrara a su sustrato físico, aun cuando su propio esfuerzo para construir ese nexo hubiera conducido a una *impasse*.[6] Lo que hay entre el psicoanálisis y la IA es la concepción de que la IA es sinónima del racionalismo, o más bien de la clase de racionalismo que está encarnado en la idea del procesamiento de la información.

Si la IA ha parecido un tanto unitaria en sus implicaciones para el pensamiento sobre las personas, es debido a que lo que muchos observadores conocen como IA es en realidad procesamiento de la información, una estrategia orientada por reglas, jerárquica, concebida para la creación de inteligencia. Pero el procesamiento de la información es sólo una parte de un cuadro más amplio.

Las dos IAs

A mediados del siglo diecinueve George Boole formalizó reglas de inferencia lógica en una forma algebraica tan suficientemente sistemática que él se sintió autorizado a llamar a su obra «Las leyes del pensamiento».[7] Por supuesto, el título de Boole llegaba más lejos que sus logros concretos, que están muy lejos de ser un modelo comprehensivo de la mente. Por empezar, las leyes de Boole necesitan un agente externo que las opere.

Las leyes de Boole son algo que una persona podría usar, pero una versión computacional de Boole inyecta vida entre sus ecuaciones. Se coloca en el sistema un operador en forma de programa de computadora. Una vez allí, se puede ver al operador y a las leyes como un modelo en funcionamiento si no de la mente, al menos de una parte de la mente.

Puede describirse una rama importante de la IA como dedicada al proyecto de Boole en forma computacional. La IA del procesamien-

to de la información otorga forma activa a proposiciones formales para crear una encarnación de la inteligencia en forma de reglas y razón. Boole formuló reglas algebraicas para la transformación de proposiciones lógicas. La moderna ciencia computacional ha ampliado lo lógico y lo proposicional, convirtiéndolos en una noción más general de lo que llama información, y ha ampliado la transformación algebraica convirtiéndola en la noción más general de procesamiento computacional. Boole reconocería cierto parentesco entre su proyecto y la forma en que Newell y Simon unieron esos dos avances en el GPS y en otros programas que establecieron las bases de la IA del procesamiento de la información.

Pero la IA no es una empresa unitaria. La computación es la sustancia de la que están hechas muchas teorías. Es verdad decir que no hay una IA sino muchas. Y ayuda mucho decir que son esencialmente dos. La primera es el procesamiento de la información, con sus raíces en la lógica, en la manipulación de proposiciones para obtener nuevas proposiciones, en la combinación de conceptos para obtener nuevos conceptos. La segunda se origina en un estilo de trabajo muy distinto, presente desde los días iniciales del campo pero con creciente influencia en tiempos recientes, al punto de constituir el foco de atención donde quiera se discuta de IA, desde los seminarios de investigación hasta los artículos populares. Esta segunda es la «IA emergente».

La IA emergente no se ha inspirado en el ordenado terreno de la lógica. Las ideas sobre la inteligencia de máquina que desarrolla tienen que ver menos con enseñar a la computadora que con dejarla que aprenda. Esta IA no sugiere que se den a la computadora reglas a seguir, sino que trata de configurar un sistema de elementos independientes dentro de la computadora, elementos de cuya interacción se espera que surja la inteligencia. Desde esta perspectiva, una regla no es algo que usted da a una computadora, sino un patrón que usted infiere cuando observa la conducta de la máquina, en forma parecida a la de la observación de la conducta de una persona. Sus imágenes sustentadoras no han sido tomadas de la lógica, sino de lo biológico.

El procesamiento de la información inyecta vida en Boole colocando un operador en su sistema, pero aquello sobre lo que opera comparte la naturaleza estática de las proposiciones de Boole. En las computadoras tradicionales, millones de unidades de información se sientan en la memoria sin hacer nada esperan-

do que el procesador central actúe sobre ellas, una a la vez. Impaciente con esta limitación, el objetivo de la IA emergente es la computación «pura». Aquí, todo el sistema es dinámico, sin distinción entre los procesadores y la información que ellos procesan. Familias de entidades similares a las neuronas, sociedades de submentes y sub-submentes antropomórficas, se encuentran en interacción simultánea. El objetivo, no menos mítico que la creación de una hebra de ADN, es la generación de un fragmento de la mente.

Las dos IAs, la orientada por reglas y la emergente, lógicas y biológicas en su estética, alimentan fantasías muy diferentes sobre la forma de construir mente a partir de la máquina. Si la IA del procesamiento de la información se halla cautivada por la imagen del ingeniero del conocimiento, hambriento de reglas, desarticulando un experto humano para encarnar los métodos del experto en algoritmos y hardware, la IA emergente está cautiva por la imagen del computador científico, despierto toda la noche observando las luces parpadeantes de una computadora, con la esperanza de que la interacción de «agentes» dentro de la máquina creará la inteligencia.

Ampliamente asociado con el espíritu y la sustancia del campo como un todo (aquí lo he llamado el estereotipo literario), el procesamiento de la información pone a la IA en una relación distante con el psicoanálisis, cuyas ideas no se traducen fácilmente en reglas y algoritmos.[8] Por cierto, ahora veremos cómo las nociones populares sobre la IA inspiradas en el procesamiento de la información sugieren que la IA es todo lo que el psicoanálisis no es. Mi tesis se sigue directamente: cuando la sustancia de la IA se expande hasta incluir no sólo la información, sino agentes interiores activos e interactivos, hay un punto de partida para un nuevo diálogo entre las culturas psicoanalíticas y de la computación.

Procesamiento de la información y psicoanálisis

El lapsus freudiano es un blanco tentador para los psicológos inclinados a encontrar mecanismos de tipo computacional en la conducta humana. Después de todo, uno comprende sólo demasiado bien las clases de errores que cometen las computadoras. ¿Qué clase de computadora cometería la clase de error que

Sigmund Freud consideró revelador de una clase de significado muy diferente? En otras palabras, ¿qué clase de computadora somos?

En *Psicopatología de la vida cotidiana*, Freud describe los lapsus verbales y toma como uno de sus ejemplos el de un presidente que inaugura una sesión parlamentaria declarándola cerrada. La interpretación freudiana de este lapsus se concentra en los sentimientos complejos que pudiera haber detrás. ¿Está el presidente angustiado ante la sesión? ¿Tiene razones para creer que nada bueno puede salir de ella? ¿Desearía más bien estar en su casa? Se supone que el lapsus nos habla sobre deseos reales. Su análisis desentraña el concepto de ambivalencia: en este caso, las emociones encontradas del presidente sobre su asistencia a la sesión.

¿Cómo podemos ver este lapsus humano como un error de procesamiento de la información? Un estudiante de computación científica del MIT no tuvo problemas para encontrar una explicación: «Se perdió un bit: el bit del signo. Puede haber sido un pico de energía eléctrica. No hay problema». Es interesante que Freud viera un problema precisamente porque *abierto* y *cerrado* son tan opuestos: sus significados opuestos otorgan significación a su sustitución. Para los estudiantes de computación científica, *abierto* y *cerrado* son conceptos próximos. En su mundo conceptual, es natural codificar conceptos opuestos como la misma raíz con diferente «bit de signo» (caliente = –frío, seco = –mojado, abierto = –cerrado). De modo que si usted piensa que la mente humana es como guardar información en la memoria de una computadora, sustituir *abierto* por *cerrado* se justifica ampliamente. Puede haber sido una pequeña falla técnica debida a algo tan trivial como un pico de energía. No necesita recurrir a la idea de ambivalencia, de deseos escondidos o de conflictos emocionales. Lo que se interpretó en términos de sentimientos cargados de sexualidad, como una ventana a los conflictos, la historia y las relaciones significantes, se convierte en un bit de información perdido o en un programa «descarrilado». Lo que el psicoanálisis interpretaría en términos de *significado*, esta psicología computacional lo vería en términos de *mecanismo*.[9]

Hay otra forma de contemplar la diferencia entre la concepción psicoanalítica del lapsus y el punto de vista del procesamiento de la información, una perspectiva que contemple la «amplitud del determinismo» en un sistema de interpretación. En tanto forma

del saber, el psicoanálisis posee una lógica que pone en juego a la totalidad de la persona para explicar todas sus acciones. Esta es la razón de que un individuo pueda usar algo tan pequeño como un error verbal para llegar a los niveles más profundos de la personalidad. Lo que pone al estudiante que dice «Puede haber sido un pico de energía eléctrica. No hay problema» en conflicto tan radical con el psicoanálisis, no es tanto que los picos de energía eléctrica sean extraños a las categorías del psicoanálisis como la idea de que un solo factor pudiera explicar un acto de lenguaje.

En la lógica tradicional, cuando usted dice «Todos los hombres son mortales; Sócrates es un hombre; por lo tanto Sócrates es mortal», su conclusión se halla determinada por dos premisas. Cambie una y obtendrá una nueva conclusión. De la misma forma, en un modelo computacional de procesamiento de la información, usted saca un bit, una pieza de información, y obtiene una nueva salida. La determinación es «angosta», como una autopista con un solo carril. El psicoanálisis utiliza una determinación «ancha». Se basa en otra clase de lógica, más parecida a la lógica que lo lleva a usted a decir que Shakespeare es un gran poeta. Encontrarse con un mal poema de Shakespeare no pone esa proposición en tela de juicio. Ni lo haría el descubrimiento de que varios de los mejores poemas de Shakespeare fueron escritos por algún otro. De este modo, aun si el presidente anunciaba que el encuentro se cerraba en el contexto de la enfermedad de su esposa, su enfermedad y el deseo de estar en casa no determinan su lapsus en ningún sentido elemental. Los fenómenos psicoanalíticos son tan «sobredeterminados» como los juicios sobre el mérito literario. Aunque abundan las imágenes populares de un libro psicoanalítico de los sueños (junto con la historia de los popularizadores que intentaron escribir uno) no hay tal cosa como un diccionario de símbolos freudianos. El significado de un sueño sólo puede descifrarse a partir de la compleja trama de las asociaciones de un soñador particular.

Pero la computación no es equivalente a la determinación estrecha del procesamiento de la información. La IA emergente construye modelos con determinación más amplia. Mientras que el procesamiento de la información otorga a conceptos como *cerrado* y *abierto* una representación simbólica concreta en una computadora, las piezas que componen la IA emergente no tienen esa clase de relación uno-a-uno con tales ideas. En la represen-

tación simbólica, el conocimiento se almacena como la copia estática de un patrón. En un sistema emergente, el patrón mismo no se almacena. Lo que se almacena son datos sobre las relaciones entre agentes de los que se espera que recreen el patrón. En esta clase de sistemas, no es posible que «un bit perdido» o «una regla cambiada» hagan diferencia en una salida. En los sistemas emergentes, las probabilidades toman el lugar de los algoritmos; la estadística toma el lugar de las reglas.

En una memoria que escribió en 1842, Lady Ada Lovelace, una amiga y mecenas de Charles Babbage, el inventor de la «máquina analítica», fue la primera persona que registró una variante de la proposición tantas veces citada respecto de que «las computadoras sólo hacen lo que usted les pide que hagan».[10] El modelo de Lovelace para pensar sobre las fuerzas y debilidades de las computadoras es paradigmático para el procesamiento de la información. Pero no se sostiene para la IA emergente. Aquí, el punto es precisamente lograr que las computadoras hagan más de lo que se les dijo que hicieran. Se ha convertido en un lugar común que la gente cite a Lovelace para defenderse de la idea de que ellos son como máquinas: «Las personas no son computadoras. No siguen reglas. Aprenden. Crecen». Pero la IA emergente se caracteriza por representaciones de las computadoras «anti-Lovelace». La IA emergente quiebra la resistencia a ver continuidad entre las computadoras y la gente describiendo computadoras que aprenden y crecen, describiendo computadoras cuya resonancia es biológica, antes que lógica.

La IA emergente y la determinación amplia

Esta resonancia biológica está ilustrada por el perceptrón, una máquina de reconocimiento de patrones diseñada a fines de la década de 1950 y un buen primer ejemplo de la IA emergente. La IA del procesamiento de la información está hecha de datos y reglas. La IA está hecha de una sustancia muy diferente, una sustancia que se capta más fácilmente en un lenguaje antropomórfico.

Imagínese que usted tiene acceso a las opiniones de mil meteorólogos estrechos de mente, cada uno de los cuales posee un método diferente y poco confiable de predicción del tiempo. Cada

uno basa su juicio en un fragmento de evidencia que puede o no estar relacionado con la predicción de la lluvia. ¿Cómo formaría un juicio? Un método de determinación estrecha en un sistema de procesamiento de la información, por ejemplo, podría ser autocrático: identificar el meteorólogo con los mejores antecedentes y votar por él. Otra estrategia, más democrática y con determinación más amplia, sería dejar que la mayoría decida. El perceptrón refina la estrategia democrática pesando cada voto con un número relacionado con el registro pasado de cada meteorólogo.

De esta forma, por ejemplo, para hacer que un perceptrón reconozca un triángulo, usted le muestra casos de triángulos y no triángulos y hace que el sistema «adivine». Sus primeras adivinanzas son al azar. Pero el perceptrón es capaz de sacar ventaja de señales que le dicen si su adivinanza es correcta o errónea para crear un sistema de votación en el que los agentes que han adivinado correctamente obtienen más peso. Los perceptrones no están programados, pero aprenden de las consecuencias de sus acciones.

En el método de determinación estrecha, usted tendría una quiebra completa si el meteorólogo escogido se volviera insano. Pero en el cerebro, el daño rara vez conduce a la quiebra total. Más a menudo produce una degradación de la performance proporcional a su extensión. En otras palabras, cuando las cosas van mal, el sistema todavía funciona, aunque no tan bien como antes. Los sistemas de procesamiento de la información pierden credibilidad como modelos de la mente porque carecen de ese rasgo; el perceptrón muestra la delicada degradación de performance típica del cerebro. Incluso con algunos meteorólogos incapacitados a bordo, el perceptrón sigue produciendo la mejor decisión posible basada en el subconjunto de actores remanentes.

En un modelo de procesamiento de la información, la conducta inteligente se sigue de reglas fijas. En el perceptrón no hay ninguna. No hay diagrama de flujo, ni un camino orientado por reglas a través del sistema. No hay tampoco correspondencias uno a uno entre información y salida. Lo que es importante no es lo que un agente sabe sino su lugar en el sistema, sus interacciones y conexiones. El perceptrón presenta un modelo de la mente como sociedad en la que la inteligencia crece a partir de la cacofonía de voces en competencia.

En un modelo de procesamiento de la información, el concepto

«lluvia» debe representarse explícitamente en el sistema. En el perceptrón la decisión «lloverá» nace de interacciones entre agentes, ninguno de los cuales posee un concepto formal de lluvia. Los perceptrones muestran la emergencia de lo que el procesamiento de la información toma como su material bruto. El procesamiento de la información comienza con símbolos formales. Los perceptrones, como el inconsciente de Freud, operan a un nivel subsimbólico y subformal. Y, lo que es más importante para esta discusión, los perceptrones descansan en las interacciones de agentes internos, objetos dentro del sistema.

La teoría del objeto es un aspecto central de la IA emergente y forma el nexo entre la IA y nuevas direcciones del pensamiento psicoanalítico. Los agentes internos en los perceptrones constituyen un puente hacia el determinismo amplio del psicoanálisis. Pero sólo se trata de una apertura. Después de los perceptrones y de los sistemas similares a los perceptrones de la década de 1960, falta todavía otra fase en el desarrollo de las ideas computacionales antes que los objetos interiores lleguen a ocupar el centro de la escena. Esta es la historia de la que ahora me ocupo, la historia de una segunda generación de IA emergente con énfasis en los objetos interiores y un nuevo camino de influencias sobre el psicoanálisis.

La IA emergente y los objetos computacionales

La atmósfera en los laboratorios de IA a comienzos de la década de 1960 era tempestuosa. La obra de Norbert Weiner, John von Neumann y Alan Turing había impulsado ondas de choque que todavía estaban frescas. Hacía muy poco que los primeros programas de procesamiento de la información que emulaban fragmentos del pensamiento humano habían producido su sorpresa. Los modelos similares al perceptrón (y había muchos de ellos, incluyendo el «Pandemonium» de Oliver Selfridge y las «redes neuronales» de Warren McCulloch) llevaban a los investigadores a descripciones de la mente artificial llenas de resonancias biológicas. Los pensamientos se concentraban en la naturaleza última de la inteligencia.

Los investigadores de la inteligencia artificial veían poca razón para un estilo humilde. Por el contrario, la IA se definía a sí misma

como una empresa de proporciones míticas: la mente creando la mente. Al hacerlo, el campo dibujaba dentro de su cultura un cierto tipo de persona, no muy diferente de la clase de persona dibujada en el círculo inicial en torno de Freud. Allí también la empresa era mítica: la comprensión racional de lo irracional. La primera generación de investigadores de la IA, con formaciones tan distintas como las matemáticas, la psicología, la economía y la física, igual que la primera generación de psicoanalistas, no había sido entrenada en «el campo» porque éste aún no existía. No había una disciplina académica. Sólo había nuevos mundos por conquistar.

A comienzos de la década de 1960, los modelos emergentes eran tanto una parte de lo que resultaba excitante para la IA como los programas de procesamiento de la información. Pero casi durante un cuarto de siglo, la IA emergente parecía haber sido abandonada. En su influencia sobre la psicología, la IA llegó casi a ser sinónimo de procesamiento de la información. Newell y Simon desarrollaron sistemas basados en reglas en su forma más pura, sistemas que simulaban el comportamiento de personas trabajando sobre una variedad de problemas lógicos. Esas simulaciones ofrecían la promesa de más, la promesa de construir una mente artificial a partir de reglas. Y si usted puede construir una mente a partir de reglas, luego puede suponerse que la mente tiene reglas en todas partes. Siguiendo esta lógica, los investigadores convirtieron los modelos de procesamiento de la información en la columna vertebral de la ciencia cognitiva.

El lenguaje del procesamiento de la información (descripciones de «búsqueda», «subrutinas», «guiones» y «gramáticas») se convirtió en la moneda común entre los psicólogos que aceptaban la idea de que los «programas de juguete», pequeñas piezas de inteligencia encarnada en la máquina, eran representativos de las grandes cosas por venir. Los programas de computadora que podían jugar al ajedrez, manipular bloques o «conversar» con camareros imaginarios en restaurantes imaginarios hacían más que modelizar pequeñas piezas del funcionamiento mental. Ellos daban sustento a la idea de que los medios utilizados para construirlos, tomados todos del paradigma del procesamiento de la información, podrían algún día capturar la esencia de la mente. Esta idea recibió apoyo adicional del éxito mundano de una clase particular de programa de procesamiento de la información, el sistema

experto. En él el científico de IA extrae reglas de decisión de un virtuoso en el campo (diagnóstico médico, por ejemplo) y las encarna en una máquina que luego hace el diagnóstico «por sí misma».

Hacia mediados de la década de 1970 la IA ya no era más marginal. Tenía sus propios programas académicos, sus propias revistas, sus propias conferencias. Se hallaba bien dotada de fondos, por su valor en el mercado y en la plaza militar. Los sistemas expertos se usaban para analizar precios, datos de perforaciones petrolíferas, material de muestras químicas. Las compañías competían para contratar graduados de IA para impulsar departamentos internos. El futuro del campo llegó a ser parte de la caliente discusión sobre la rivalidad industrial japonés-norteamericana.

Ahora la IA podía prometer una clase más tradicional de carrera, lo mismo que la medicalización del psicoanálisis había allanado el camino para convertirlo en una especialidad psiquiátrica profesionalizada. Tanto en el psicoanálisis como en la IA, las carreras tradicionales significaban nuevas presiones para comprometerse en la clase de tareas que prometían resultados visibles. En el psicoanálisis la presión era «curar», trabajar en problemas educacionales, hacer «psicoanálisis aplicado». En la investigación en IA el péndulo se movió desde lo que había sido más mítico en las décadas de 1950 y 1960 a lo que la gente «sabía cómo hacer»: recopilar reglas y codificarlas en programas de computadora.

Pero incluso a medida que el modelo del procesamiento de la información alcanzaba una casi hegemonía a fines de la década de 1970, se estaban desarrollando las condiciones para algo muy diferente. Primero que nada, había habido un importante progreso técnico. Los computadores científicos hacía mucho que protestaban contra las limitaciones de la computadora de von Neumann, en las que un procesador puede manipular los datos pasivos en un millón de células de memoria. Siempre había sido obvio que, en principio, la distinción entre procesador y memoria podía abolirse haciendo que cada célula de la computadora fuera un procesador activo. Hacerlo, sin embargo, había sido prohibitivamente caro. Pero ahora, proyectos como la Máquina de Conexión eran suficientemente realistas como para ser subsidiados. En éste, el plan es tener un millón de microprocesadores juntos para hacer una

computadora cuya memoria y fuerza computacional estuvieran ampliamente distribuidas. Ya no habría más un operador y material pasivo sobre el cual operar. La computación estaba «despertando del sueño booleano».[11]

Junto con el hardware, que presentaba nuevas posibilidades, había nuevas ideas sobre cómo programarlo. El desarrollo de metodologías de programación con nombres tan sugestivos como «traspaso de mensajes» y «modelos de actor» creó el contexto para pensar en agentes computacionales en comunicación. Los programas de computación estándar son listas de instrucciones en forma de imperativos: «agregar estos números», «poner el resultado en la memoria», «obtener el contenido de esta dirección de memoria». Los programas de inteligencia artificial en LISP o Prolog operan sobre datos más abstractos, pero aún consisten en instrucciones para manipular información. El primer cuarto de siglo del desarrollo de la programación se basaba en un lenguaje de procesamiento para describir cómo pasar información de un lugar a otro. Pero los investigadores sentían ahora la necesidad de tratar con una nueva clase de suceso: no el *traspaso* de algo, sino la *creación* de algo. Por una coincidencia que llega a ser enormemente sugestiva a los fines de la presente discusión, los computadores científicos llamaron a lo que es hasta hoy su respuesta más prominente «programación orientada al objeto».

Si usted quiere simular una fila de clientes en un mostrador de una oficina de correos (para averiguar, por ejemplo, cuál sería el promedio de duración de la espera si el número de empleados se redujera a uno), usted escribe un programa que crea un objeto interno que se «comporta como» una persona en una fila en la oficina de correos. Esta avanza cuando lo hace la persona que está adelante; sabe cuándo llega al mostrador y procede a desarrollar su transacción. El contraste entre esta estrategia orientada al objeto y las estrategias tradicionales de programación es dramático. Un programador tradicional en FORTRAN asignaría *x*s e *y*s a las propiedades de los clientes y escribiría código de computación para manipular las variables. La programación orientada al objeto se refiere directamente a los objetos interiores que representan a los clientes en la fila: las *x* y las *y* no aparecen.

En la programación orientada al objeto, el programador construye nuevos objetos que, una vez creados, se pueden «dejar en libertad» para que interactúen de acuerdo con la naturaleza con

que se los ha dotado. El programador no especifica qué es lo que harán concretamente los objetos, sino más bien «quiénes son».

Si los diagramas de flujo captan parte del «sentimiento» de un programa de procesamiento de información, parte del «sentimiento» de la programación orientada al objeto es captada por las imágenes de las carpetas de archivo, las tijeras y el cesto de desperdicios que aparecen en las pantallas de las computadoras con interfaces gráficas. Los iconos son el reflejo superficial de una filosofía de programación en la que las computadoras se piensan como «teatros electrónicos de marionetas», y en la que «no hay limitaciones importantes a las clases de obras que se puedan desarrollar en sus pantallas, ni al rango de vestimentas o papeles que los actores puedan asumir».[12] Para los matemáticos, las manipulaciones algebraicas en los programas tradicionales poseen una realidad apremiante. Para los no matemáticos, la estrategia de programación orientada al objeto posee un atractivo más directo, el atractivo de los actores en un escenario.

Hacia comienzos de la década de 1980 la coexistencia del nuevo hardware paralelo y las nuevas ideas sobre objetos en la programación definieron la escena para que el péndulo se apartara del procesamiento de la información. El comienzo de la década presenció la primera de una creciente serie de artículos de orígenes muy diversos: ingenieros ansiosos de construir nuevas máquinas paralelas, computadores científicos ansiosos de tratar nuevas ideas matemáticas que pudieran orientar nuevos esfuerzos en la programación en paralelo, psicólogos en busca de nuevos modelos que poseyeran una resonancia biológica (y más concretamente, neurológica). La IA emergente no había muerto, sino que estaba soterrada. Resurgió con una venganza en mente y con un nuevo nombre: «conexionismo». Una vez más, no habría distinción entre el procesador y lo que se procesaba. No había un conjunto de operaciones especificadas. Sólo habría comunidades de agentes en interacción recíproca directa.

Pero los proponentes de la nueva teoría del conexionismo iban más allá de las etapas iniciales de la IA emergente en los pasos que deseaban dar para alejarse de Boole. Por ejemplo, el perceptrón no podía generar por sí mismo nuevos objetos o elucidar la forma en que podrían emerger conceptos nuevos. Sus agentes estaban programados por un humano que actuaba desde el exterior. Los conexionistas de hoy esperan ir más lejos poniendo en conjunto

máquinas paralelas, la maduración de ideas sobre su programación y, lo que es más importante, un nuevo sentido sobre el problema central que enfrenta el campo, algo que rara vez se había formulado durante la década de 1960: ¿Cómo se crean los objetos?

Objetos psicoanalíticos y modelos de la sociedad

En el foco de atención de los objetos interiores y su emergencia e interacción, la IA comparte preocupaciones que son centrales para la teoría psicoanalítica contemporánea. Como ha sido el caso en la IA, el desarrollo de una nueva teoría del objeto psicoanalítico es un desarrollo tardío en el campo. No fue allí donde comenzó la teoría.

La teoría psicoanalítica primitiva estaba construida en torno del concepto de pulsión, la demanda generada por el cuerpo, que proporciona la energía y los objetivos para toda la actividad mental. Pero después, cuando Freud volvió su atención hacia la relación entre ego y el mundo exterior, la significación y la estructura de estas relaciones ya no podían estructurarse según la teoría de la pulsión.

Hacia 1917 Freud comenzó a formular un lenguaje para tratar estas cuestiones. Describió un proceso mediante el cual la gente forma «objetos» interiores. En *Duelo y melancolía*, Freud argumentó que los sufrimientos del melancólico proceden de los reproches mutuos entre el sujeto y un padre internalizado con el que el sujeto se identifica. En este artículo Freud describe la «introyección» de personas (en jerga psicoanalítica, objetos, y en este caso el padre) como parte de una patología, pero luego llegó a la conclusión de que este proceso es parte del desarrollo normal. Por cierto, éste es el mecanismo para el desarrollo del superego, la introyección del padre ideal.

De acuerdo con Freud, internalizamos objetos porque nuestros instintos nos impelen a hacerlo. En su obra, el concepto de objetos interiores necesitaba coexistir con el andamiaje de la teoría de las pulsiones. Pero muchos teóricos psicoanalíticos que lo siguieron estaban menos comprometidos con el modelo de la pulsión. Ellos ampliaron el espectro de lo que Freud quería decir con «relaciones objetales», al punto que ahora pensamos que ellos constituyen

una escuela distintiva. La teoría freudiana clásica posee muchos conceptos solapados para describir objetos internos: huellas de memoria, representaciones mentales, introyecciones, identificaciones y la idea de estructuras internas tales como el superego. La estrategia de las relaciones objetales es menos específica sobre nuestros contenidos. Ella describe una sociedad de agentes interiores o «micromentes», «suborganizaciones inconscientes del ego capaces de generar significado y experiencia, es decir, capaces de pensamiento, sentimiento y percepción».[13] Las relaciones con las personas, introyectadas como entidades internas, son las piezas de construcción fundamentales de la vida mental.

Mientras Freud concentraba su atención en un solo objeto internalizado, el superego, los teóricos de las relaciones objetales describen un mundo interior ricamente poblado. La psicoanalista Melanie Klein fue tan lejos como para caracterizar la gente que los niños traen dentro (así como las representaciones de las partes del cuerpo) como poseedora de rasgos psicológicos, personalidades. Se las podía ver como amantes, odiosas, voraces, envidiosas. El psicoanalista W. R. D. Fairbairn rearticuló los motores básicos freudianos del desarrollo de la personalidad en términos de relaciones objetales. Para Fairbairn, el organismo humano no es impulsado por el principio freudiano del placer, el deseo de reducir la tensión pulsional, sino más bien por su necesidad de establecer relaciones. Esto constituye una profunda reformulación de la perspectiva psicoanalítica del sujeto: la gente no busca fundamentalmente placer, sino objetos.

El lenguaje que el psicoanálisis necesita para hablar sobre objetos (cómo se forman, cómo interactúan) es muy diferente del lenguaje que necesita para hablar sobre pulsiones. En su «Proyecto para una Psicología Científica», Freud trató de usar términos de tipo informacional derivados de la descripción del arco reflejo (la fibra sensorial transporta información al cerebro, la fibra motora transporta información al músculo) para hablar sobre la memoria, los instintos y el flujo de la energía psíquica. Pero la metáfora de la información fracasa completamente cuando se la usa para hablar de objetos internos. Así como es en programación orientada al objeto, así es en el psicoanálisis. Cuando uno habla de objetos, las metáforas naturales tienen que ver con hacer algo, y no con transportar algo.

En la teoría psicoanalítica clásica, unas pocas y poderosas

estructuras internas (el superego, por ejemplo) actúan sobre los recuerdos, pensamientos y deseos. La teoría de las relaciones objetales postula un sistema dinámico en el que la distinción entre procesador y procesado se viene abajo. El paralelismo con la computación es claro: en ambos casos hay un movimiento que se aparta de una situación en las que unas pocas estructuras internas actúan sobre una sustancia pasiva. Fairbairn reemplazó las dicotomías freudianas de ego y ello, estructura y energía, con agencias independientes dentro de la mente que piensan, desean y generan significado en interacción recíproca, en forma muy parecida a la de la definición de agentes autónomos dentro de la computadora en IA.

El desarrollo de la teoría de las relaciones objetales ha llevado al psicoanálisis a preguntarse si su lealtad a Freud depende de aceptar su modelo de la pulsión. Algunos han tratado de preservar el lenguaje pulsional original de Freud, pero usándolo de una manera que incorpora nuevos énfasis en relaciones objetales, por ejemplo, asignando un papel a los objetos en relación con la descarga de la pulsión: pueden inhibir, descargar, facilitar o servir de blanco a la pulsión. Pero esta reelaboración del vocabulario es menos una solución que un intento de paliar el problema. Sólo funciona si los objetos internos no poseen propiedades elaboradas o si su creación se concibe como un suceso ocasional.

Pero cuando los objetos se vuelven centrales para la comprensión de la psique, hay mayor presión para alejarse de la teoría de la pulsión. Aunque la teoría de la pulsión se ha tornado cada vez más sofisticada y abierta a la discusión de objetos interiores, la escisión entre una estrategia pulsional y una estrategia de relaciones objetales es en la actualidad una división central en el psicoanálisis.[14] La división es paralela a la que se da entre el procesamiento de la información y la IA emergente. Para usar el lenguaje de Thomas Kuhn, los teóricos de las relaciones objetales están diciendo que el psicoanálisis ya no puede proceder como una «ciencia normal»,[15] que crece mediante la asimilación de nuevos datos en la vieja teoría. Para ellos, las relaciones objetales constituyen un cambio de paradigma dentro del psicoanálisis, lo mismo que la hipótesis de la emergencia —que la inteligencia se desarrolla a partir de la interacción de múltiples agentes (no es lo que usted sabe, sino quién es usted)— representa un cambio de paradigma en IA.

Los teóricos de la inteligencia artificial Marvin Minsky y Seymour Papert han construido un modelo computacional que evoca la teoría de Fairbairn de las relaciones objetales. Sus teorías toman la mente como una sociedad de agentes en interacción. Estos agentes están antropomorfizados, se discuten en los términos que uno usualmente reserva para las personas totales, pero no poseen la complejidad de la gente. Por cierto, su modelo se basa (como el concepto del perceptrón) en el hecho de que esos agentes son «estúpidos». Cada uno sabe una cosa y solamente una cosa. Y, al igual que los «agentes votantes» del perceptrón, su estrechez de visión los lleva a opiniones muy discrepantes. La estructura compleja de la conducta, la emoción o el pensamiento emergen del conflicto de sus visiones opuestas.

La presentación más elaborada de esta teoría, el libro de Minsky *The Society of Mind*, describe una vasta matriz de agentes: agentes censores, agentes de reconocimiento y agentes de la ira, para mencionar sólo unos pocos.[16] No es sorprendente que Minsky reconozca en Freud, quien escribió extensivamente sobre agentes censores, a un colega en la modelización de la «sociedad». Es más sorprendente que Minsky vea a los actores como actores claves, no sólo para modelizar el pensamiento humano, sino para hacer máquinas inteligentes.

La idea de Minsky de un censor es un ejemplo dramático de la resonancia que se está desarrollando entre la IA emergente y el psicoanálisis. El censor de Freud protege a la gente de los pensamientos dolorosos. La extensión de esta idea al funcionamiento cognitivo y a los «pensamientos» en una máquina no depende del supuesto de que los agentes o el sistema como un todo sienten dolor. Para funcionar en forma coherente, de acuerdo con Minsky, un sistema inteligente debe desarrollar una cierta falta de atención a sus voces agentes en contradicción. La formulación de Minsky establece que no puede haber inteligencia, artificial o lo que sea, sin represión. Allen Newell ha hablado de la necesidad de censores en los sistemas grandes y complejos de procesamiento de la información. Pero con reglas claras e inambiguas estipuladas de antemano, una computadora procesadora de información también se las puede arreglar sin ellos. Los censores pueden llegar a ser prácticos, pero no son necesidades teoréticas del paradigma del procesamiento de la información. En el caso de la teoría de la sociedad, sin embargo, los censores son intrínsecos. Dado que no

puede haber inteligencia sin contradicción y conflicto, sólo la presencia de censores permite que la inteligencia emerja.

En esto, la teoría de la sociedad y la teoría freudiana comparten un punto importante. Freud no «descubrió» el inconsciente. Su contribución fue la elaboración de un inconsciente *dinámico*. Lo que es inconsciente no es simplemente olvidado, viejo o irrelevante para el funcionamiento actual. Es reprimido. Hay fuerzas poderosas que lo mantienen encubierto, y por una buena razón. En forma parecida, para Minsky, lo que es reprimido en la máquina computacional y lo que él ha llamado «la máquina de carne» humana *necesitan* ser reprimidos.

Freud escribió sobre los efectos de la represión de las experiencias atemorizadoras y cargadas de emociones. Minsky extiende las ideas de Freud al dominio cognitivo. «La mente de un niño pensante [no necesita ninguna] cuando alguna paradoja la envuelve y la aprisiona en un ciclón.» La paradoja, argumenta Minsky, es tan peligrosa como la escena primaria. El niño sabe que está en presencia de una amenaza cuando se le pide que trace los límites inexistentes entre los océanos y los mares o que considere cuestiones sobre el huevo y la gallina, sobre lo que sucedía antes del comienzo de los tiempos y sobre la localización de los límites del espacio. Minsky añade: «¿Y qué hay de frases tales como *'Esta afirmación es falsa'* , que puede empujar a la mente a girar como un trompo? No sé de nadie que recuerde esos incidentes como si fueran amenazantes. Pero entonces, como Freud podría decir, este mismo hecho debe ser un indicio de que el área está sujeta a censura».[17]

Minsky siente que las nociones de «represión cognitiva» y el «inconsciente cognitivo» nos permitirán ir más allá de Freud. Utiliza como ejemplo la elaboración freudiana sobre el chiste. El trabajo de Freud de 1905 sobre el chiste explica que los censores interiores sirven como barreras contra los pensamientos prohibidos. La mayoría de los chistes está diseñada para engañar a los censores. Es una forma de disfrutar de un deseo prohibido. Esta es la razón por la que muchos chistes involucran tabúes concernientes a la crueldad y la sexualidad. Pero a Freud le preocupaba que esta teoría no diera cuenta fácilmente de los «chistes sin sentido». Una de las hipótesis de Freud sobre el poder de los chistes sin sentido era que el sinsentido refleja «un deseo por retornar a una niñez despreocupada, cuando se nos permitía

pensar sin la compulsión de ser lógicos». La idea del inconsciente cognitivo da sustento a esta perspectiva: la paradoja y el sinsentido necesitan reprimirse en el proceso de desarrollo de la inteligencia emergente, sea en las máquinas o en las personas. Los resultados absurdos del pensamiento son tabú, tan amenazadores como el sexo. Los censores trabajan igual de duro para suprimirlos; ellos no tienen inocencia.

Subversión y normalización

A despecho de sus diferencias, el psicoanálisis y la IA siempre han compartido afinidades teoréticas, entre ellas, como hemos visto, el desafío a la idea del actor autónomo e intencional, la necesidad de la autorreferencia en la construcción de la teoría y la necesidad de objetos tales como censores para tratar con los conflictos internos. Pero la afinidad se tornó algo más fuerte cuando el conglomerado de cuestiones sobre los objetos pasó a ocupar para ambos el centro de la escena. Esta nueva orientación ha hecho que los viejos elementos en común sean más comunes: las teorías del agente en IA subrayan preocupaciones teoréticas que reproducen las preocupaciones del psicoanálisis. Estas incluyen el conflicto, la consistencia interna y, quizá más dramáticamente, la subversión del sujeto, el sujeto «descentrado».

Aunque tanto el psicoanálisis como la IA siempre han desafiado al «Yo» actor, ambos poseen variantes teoréticas que subrayan este desafío más que otros. El inconsciente freudiano socava la idea del sujeto unificado, pero muchos de los seguidores de Freud procuraron restaurar el sentido de la existencia de un ejecutivo mental concentrando su atención en el ego, esa parte del sujeto dividido de Freud que se volcaba hacia la realidad exterior. Estos «psicólogos del ego» comenzaron a hablar de él como un agente capaz de integrar la psique. Anna Freud escribió sobre su poderosa artillería, los mecanismos de defensa. Heinz Hartmann argumentó que el ego poseía un aspecto que no estaba amarrado a las neurosis del individuo, una zona «libre de conflictos». Hartmann escribió sobre este aspecto desinhibido del ego como si fuera libre de actuar y de elegir, independientemente de toda coacción. Casi parecía el asiento para una versión renacida del libre albedrío, el lugar de la responsabilidad moral. El historiador

intelectual Russell Jacoby, escritor del ego renacido de la psicología, el «Yo» autónomo, llegó a llamarlo «el olvidado del psicoanálisis».[18]

En su forma subversiva, que escinde el ego y socava al sujeto, el psicoanálisis es difícil de digerir. Vuela sobre el rostro de la comprensión de sentido común. Es una ciencia subversiva. La psicología del ego la normaliza. Toma lo que es más subversivo —el sujeto descentrado— y lo suaviza. La psicología del ego presenta la versión del inconsciente más aceptable para la conciencia.

Este patrón de respuesta normalizadora es común a todas las ciencias subversivas de la mente, incluyendo, por supuesto, a las computacionales. Hemos visto que la idea misma de la IA ponía en cuestión al sujeto bajo la forma del programa. Pero la IA, también, poseía variantes que suavizaban su mensaje sobre el descentramiento. Por ejemplo, si usted reduce la IA a la idea de los sistemas expertos, hay un paso muy pequeño de allí a pensar que el sistema experto es un recurso al que algún ejecutivo central no claramente especificado puede invocar. Cuando usted comienza con la idea de que una computadora puede tener esos ejecutivos y esos recursos, la idea de que un humano también los tiene se sigue directamente. Eso hace que el modelo de la mente de la IA parezca menos amenazador, porque lo que se necesita pensar como algo computacional y orientado por reglas no es mi «Yo», sino mi «expertez» en un dominio limitado, por ejemplo, la parte de mí que juega al ajedrez. El sujeto se convierte en el ejecutivo que supervisa al experto. De modo que hay versiones, tanto del psicoanálisis como de la IA que le sacan la mecha al principio subversivo del descentramiento, restringiendo su papel a la explicación de partes de la mente y evitando así el riesgo de disolver el todo.

La estrategia de neutralizar la teoría subversiva es menos viable en el caso de las teorías del agente y de los objetos, que son más agresivas en su negación del sujeto unificado. Por cierto, estas teorías se definen a sí mismas a través de esa negación. Ponen al psicoanálisis y a la IA en una relación nueva y más estrecha entre ellos y con otros movimientos intelectuales que «desconstruyen» el sujeto humanista.

La fuerza y la debilidad de las teorías del objeto son las mismas en el psicoanálisis y en la IA: la fuerza es un marco conceptual que ofrece ricas posibilidades para los modelos de los procesos inte-

ractivos; la debilidad es que ese marco puede ser *demasiado* rico. Los objetos postulados pueden ser demasiado poderosos: explican la mente postulando muchas mentes dentro de ella. La teoría de los objetos confronta ambos campos con el problema de la regresión infinita. Hay algo profundamente insatisfactorio en una teoría que no puede ir más allá de asumir un homúnculo dentro del humano, pues ¿cómo explicamos entonces el homúnculo interior sin postular otro homúnculo dentro de él, y así sucesivamente?

Los teóricos psicoanalíticos luchan con esta cuestión. Dentro del campo, gran parte de la crítica hacia los objetos interiores demasiado poderosos ha tomado como blanco la obra de Melanie Klein. Por ejemplo, el psicoanalista Roy Schafer ha argumentado que Klein y la escuela inglesa de relaciones objetales ha llevado la reificación implícita en la metapsicología freudiana a un «extremo grotesco»: «Una multitud de mentes se introduce en un solo aparato psíquico. La persona se percibe como contenedora de innumerables microorganizaciones independientes, que son también microdinamismos.»[19] Esencialmente, los críticos de Klein sienten que su idea de las «figuras interiores idealizadas que protegen al ego contra las ideas aterradoras es tanto como proponer que hay 'demonios' internos amistosos y hostiles operando en la mente».[20]

Los kleinianos responden que esas figuras interiores no son demonios sino fantasías y pensamientos inconscientes. Son las ideas que tenemos sobre lo que contenemos.[21] Pero es dudoso que su respuesta zanje la cuestión. El psicoanalista Thomas Ogden plantea el problema de plano: ¿Cómo puede ser que los pensamientos se comporten como agentes?

Si los objetos interiores son pensamientos... luego no pueden pensar, percibir o sentir por sí mismos, ni pueden proteger ni atacar al ego. Incluso en la actualidad, los teóricos kleinianos no son capaces de desenredarse de Escila de la demonología y de Caribdis de mezclar niveles incompatibles de abstracción (es decir, agencias y pensamientos activos).[22]

En computación científica, el conexionismo no ha resuelto el problema de dar cuenta de los objetos (qué son y cómo llegan a serlo). El conexionismo simplemente postula los agentes interio-

res que requiere, por lo cual el científico de IA Terry Winograd ha llegado a decir que parte de su atractivo es que «posee un porcentaje más alto de racionalización del deseo».[23] Pero el problema de la regresión infinita (dar cuenta de las entidades que dan cuenta del pensamiento) posee un alcance distinto en IA que en psicoanálisis, porque los computadores científicos están acostumbrados a descansar en una forma controlada de razonamiento circular (la «recursión») como una poderosa herramienta técnica.

La mayoría de nosotros aprendió en la escuela a definir x elevada a la potencia n como x multiplicada por sí misma n veces. La potenciación se define en términos de la multiplicación. Los computadores científicos prefieren definir x elevada a la potencia n como la potencia n-1 multiplicada por x. La potenciación se define en términos de la potenciación. A partir de ejemplos tan sencillos, compartidos por las matemáticas precomputacionales, la computación científica ha construido una cultura matemática que descansa fuertemente en la definición de las cosas en términos de ellas mismas.[24]

Para el psicoanalista Ogden, la idea de que un pensamiento pueda pensar es impensable. La IA del procesamiento de la información también divide el pensamiento del acto de pensar. Lo que está más cerca de este último es su procesamiento. Pero la IA rompe con esta distinción. Toma la idea de recursión y la transforma en una estética envolvente. Para decirlo en forma más tajante, la IA emergente proporciona una salida al problema de la regresión infinita, redefiniendo el problema como una de las fuentes de su fuerza. Al servirse de la recursión como de una estética científica, la IA encuentra una salida del agujero teorético. Podría ser que la recursión ofrezca al psicoanálisis una posibilidad semejante.

Lo que la memoria computacional fue para el nacimiento de la ciencia cognitiva en la década de 1950, la recursión podría ser para los estudios psicoanalíticos de la década de 1990. Podemos imaginar que los computadores científicos procurarán sustentar el psicoanálisis kleiniano construyendo un detallado modelo computacional de los objetos kleinianos. Pero también podemos imaginar que teóricos psicoanalíticos computacionalmente refinados encontrarán en la idea recursiva de que los pensamientos pueden pensar más una virtud placentera que un vicio devastador. Podemos imaginarnos que los psicoanalistas teóricos verán

ideas recursivas en su trabajo como una fuente de legitimación antes que como un signo de debilidad.

No se puede realizar una predicción simple sobre la forma en que la recursión ayudará al psicoanálisis a tratar con la regresión infinita de las teorías del objeto, pero parece probable que la clase de influencia que ha de esperarse será que el psicoanálisis resulte cada vez más permeable a la recursión como mito sustentador. Esto haría que lo mismo que ha sido esgrimido como crítica se convierta en una forma de dar sustento a la teoría. En el espíritu de la concepción de George Miller sobre la memoria computacional y el conductismo, el psicoanálisis puede encontrar que resulta embarazoso negar a los pensamientos humanos la facultad de pensar, cuando los «pensamientos computacionales» se supone que lo hacen.

Si la cuestión central en la teoría psicoanalítica actual volviera a ser la naturaleza del «instinto de muerte», serían de escasa utilidad las teorías relativas a una máquina que nunca pasó por el trance de nacer. Pero en la medida en que las preocupaciones teóricas del psicoanálisis tienen que ver con la estructura y funcionamiento de objetos internos, el psicoanálisis se está moviendo hacia la IA, hasta el punto en que el camino para un diálogo productivo ya parece abierto.

Cuando el diálogo comience, la influencia de la IA sobre el psicoanálisis no dependerá necesariamente del hecho de que la IA ofrezca consejo técnico, sino de si puede ofrecer soporte moral a los teóricos del objeto psicoanalítico sitiados en su debate. ¿Puede servir como mito sustentador? La influencia de la IA sobre la psicología, psicoanalítica u otra, se relaciona no sólo con la resolución de problemas técnicos sino con el crecimiento de las culturas psicológicas.

Cultura psicoanalítica y cultura computacional

Las culturas psicológicas no sólo existen en el mundo de los profesionales. La inteligencia artificial y el psicoanálisis constituyen el contexto en el que psicólogos profesionales y psicólogos amateurs piensan sobre el pensamiento. Desde una perspectiva sociológica sobre esta cultura psicológica más amplia, las teorías del objeto hacen que las ideas en IA y en psicoanálisis sean más

«apropiadas», más fáciles de asimilar por las personas como forma de pensar sobre sí mismas que las teorías de la información o de las pulsiones. En otras palabras, las teorías objetales dan al psicoanálisis y a la IA una presencia mayor como filosofías en la vida cotidiana. Los densos textos de Fairbairn y la teoría matemática del conexionismo pueden no ser más accesibles a los legos que los artículos técnicos sobre el procesamiento de la información o sobre pulsiones psicoanalíticas. Pero cuando las teorías orientadas al objeto se popularizan y se mueven hacia la cultura general, adquieren un atractivo especial. Las ideas sobre los objetos y los agentes son más concretas que las ideas sobre pulsiones o diagramas de flujo. Son seductoras porque es fácil «jugar» con ellas. Hablan de un problema común. Todos tenemos la experiencia de no sentirnos por completo «al unísono» con nosotros mismos: las voces interiores nos ofrecen consejos conflictivos, reaseguros y castigos. Estas experiencias se traducen fácil y satisfactoriamente al drama de los objetos interiores.

Las ideas freudianas sobre los lapsus verbales se hicieron muy conocidas y ganaron una amplia aceptación por razones que tenían muy poco que ver con la evaluación positiva de su validez científica. Los lapsus verbales freudianos se hicieron parte de la más amplia cultura psicológica porque resulta fácil jugar con lo que puede estar escondido detrás de ellas. Los lapsus son casi ideas tangibles. Son manipulables. Los lapsus son atractivos como objetos en los que pensar. Usted puede analizar sus lapsus y los de sus amigos. La teoría de los lapsus proporciona a las ideas psicoanalíticas una forma de convertirse en parte de la vida cotidiana. Ayudan a que la teoría psicoanalítica sea apropiable.

Una perspectiva freudiana sobre la apropiabilidad de las ideas psicoanalíticas puede ir más allá de sugerir que la teoría de los lapsus significantes es atractiva porque nos pone en contacto inmediato con el tabú. Tenemos miedo de los aspectos sexuales y agresivos de nuestra naturaleza, pero también deseamos estar en contacto con ellos. Las ideas psicoanalíticas nos proporcionan una forma de jugar con lo prohibido. De la misma manera, tenemos miedo de pensar en nosotros mismos como si fuéramos máquinas, pero buscamos formas de reconocer esta parte muy real, aunque perturbadora, de nuestra experiencia. Jugar con la IA, con la idea de la mente como computadora, lo hace posible. Ahora bien, jugar con las teorías psicoanalíticas y computaciona-

les de los objetos y los agentes nos permite ir todavía más lejos. La idea de los agentes nos proporciona una forma de reconocer la experiencia de la fragmentación. Los sesgos racionales en nuestra cultura presentan consistencia y coherencia como si fueran naturales, pero los sentimientos de fragmentación abundan. Por cierto, se ha argumentado que constituyen una enfermedad cultural contemporánea.[25] Las teorías del psicoanálisis y la IA que hablan simple y dramáticamente a la experiencia del sujeto escindido poseen una fuerza particular.

En el pasado, la cultura de la computadora y la cultura psicoanalítica han estado separadas. En general, las ideas psicoanalíticas para pensar el sujeto congeniaban con las personas que tenían escaso contacto con las ideas computacionales. Cuando los miembros de la cultura psicoanalítica se encontraban con modelos computacionales de la mente, lo más probable es que fueran modelos de procesamiento de la información que parecían fuera de tono con la mirada psicoanalítica. Estos modelos describían secuencias, no asociaciones, y su modelo de determinación era estrecho, más que amplio. Pero cada vez con más frecuencia las ideas computacionales reportadas en la prensa popular y en la prensa académica no son sobre reglas e información sino sobre agentes, conexiones y sociedades de la mente. Estas nuevas metáforas poseen una estética biológica: son la clase de cosas que pueden estar pasando en el cerebro. Sugieren una determinación amplia y una represión dinámica. Describen un sistema en conflicto. Y lo que es más importante, son consonantes con las ideas psicoanalíticas que actualmente se difunden, ideas no sobre pulsiones y sus vicisitudes, sino sobre objetos y su interacción.

Cuando la presencia de la computadora legitimó la idea de la memoria, estaba reforzando una idea sobre psicología que era anterior a la computación. Pero las ideas sobre la recursión y los agentes no son precomputacionales. ¿Nos atreveremos a especular qué pasará entre la computación y nuestra cultura psicológica si la IA encuentra una voz finalmente divorciada de lo que quedaba estático, clavado en la lógica, y si el psicoanálisis encuentra una voz finalmente divorciada de las cuestiones de la teoría pulsional del siglo diecinueve?

Notas

[1] Muchas de las ideas en este ensayo surgieron de una serie de conversaciones con Seymour Papert, un colaborador en mi noción del papel de los mitos sustentatorios en la sociología de la ciencia de la mente.

[2] Citado en Jonathan Miller, *States of Mind* (Nueva York: Pantheon, 1983), 23.

[3] Ibid.

[4] Citado en Sherry Turkle, *The Second Self: Computers and the Human Spirit* (Nueva York: Simon y Schuster, 1984), 256.

[5] Citado en Jeremy Bernstein, *Science Observed* (Nueva York: Basic Books, 1982), 110-11.

[6] Sigmund Freud, «Project for a Scientific Psychology», *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, vol. 1, traducción y edición de James Strachey (Londres: Hogarth Press, 1960).

[7] George Boole, *The Laws of Thought*, vol. 2 de sus *Collected Works* (La Salle, Ill.: Open Court Publishing Company, 1952).

[8] Un esfuerzo sugestivo para construir algoritmos psicoanalíticos fue llevado a cabo por el psicoanalista Jacques Lacan en su teoría de los *mathèmes*. La fuerza de esta idea deriva de su esfuerzo para legitimar la sistematicidad y una relación más estrecha con la ciencia en los estudios psicoanalíticos. Véase Sherry Turkle, *Psychoanalytic Politics: Freud's French Revolution* (Nueva York: Basic Books, 1978).

[9] Para un ejemplo de una perspectiva de lo freudiano en términos de procesamiento de la información, véase Donald Norman, «Post-Freudian Slips», *Psychology Today*, abril de 1980:41-44 y sigs.; Norman, *Slips of the Mind and an Outline of a Theory of Action* (San Diego: Center for Human Information Processing, University of California, noviembre de 1979); y Norman, «Categorization of Action Slips», *Psychological Review* 88 (enero de 1981):1-15.

[10] Lovelace lo escribió de esta manera: «La Máquina analítica no tiene en absoluto pretensiones de originar nada. Ella puede hacer cualquier cosa que nosotros sepamos ordenarle que ejecute».

[11] Esta frase ha sido tomada en préstamo de Douglas R. Hofstadter, quien discute sobre computación y la estética booleana en «Waking Up From the Boolean Dream, or Subcognition as Computation», en *Metamagical Themas: Questing for the Essence of Mind and Pattern* (Nueva York: Basic Books, 1985).

[12] Alan Kay, «Software's Second Act», *Science* 85 (noviembre de 1985):122.

[13] Thomas H. Ogden «The Concept of Internal Object Relations», *The International Journal of Psycho-Analysis* 64 (1983):227.

[14] Véase Jay R. Greenberg y Stephen A. Mitchell, *Object Relations in Psychoanalytic Theory* (Cambridge: Harvard University Press, 1983).

[15] Thomas Kuhn, *The Structure of Scientific Revolutions*, 2ª edición (Chicago: University of Chicago Press, 1970).

[16] Marvin Minsky, *The Society of Mind* (Nueva York: Simon y Schuster, 1987).

[17] Ibid., 183. El trabajo de campo con niños y computadoras es rico en ejemplos de la clase de temor que Minsky espera. Por ejemplo, se reporta un incidente en el que el temor fue suscitado por el primer contacto con la recursividad en Sherry Turkle, *The Second Self*. Las entrevistas con adultos sobre experiencias tempranas también revelan muchos de esos recuerdos: el temor a los prismas, de los espejos que reflejan espejos, el miedo a preguntas tales como «¿Cuán lejos están las estrellas?».

[18] Russell Jacoby, *Social Amnesia: A Critique of Contemporary Psychology from Adler to Laing* (Boston: Beacon Press, 1975).

[19] Roy Schafer, *A New Language for Psychoanalysis* (New Haven: Yale University Press, 1976), 3; y Schafer, *Aspects of Internalization* (Nueva York: International University Press, 1968), 62.

[20] Ogden, «Internal Object Relations», 229.

[21] Hannah Segal, *Introduction to the Work of Melanie Klein* (Londres: Hogarth Press, 1978).

[22] Ogden, «Internal Object Relations», 230.

[23] *Science* 86 (mayo de 1986):27.

[24] La estética computacional del pensamiento recursivo ha sido expresada en forma poética y accesible por Douglas R. Hofstadter, quien presenta los fenómenos recursivos como una de las fuentes de la fuerza de la música de Bach, del arte de Escher y de las matemáticas de Gödel. Véase *Gödel, Escher, Bach: An Eternal Golden Braid* (Nueva York: Basic Books, 1978).

[25] Véase, por ejemplo, Christopher Lasch, *The Culture of Narcissism* (Nueva York: Norton, 1979).

12
Mucho ruido por muy poco

Hilary Putnam

La cuestión que deseo contemplar es ésta: ¿Nos ha enseñado la inteligencia artificial algo de importancia acerca de la mente? Me inclino a pensar que la respuesta es no. También me inclino a preguntarme: ¿Sobre qué es todo este alboroto? Por supuesto, es posible que la IA nos enseñe algo importante sobre la forma en que pensamos, pero ¿por qué estamos ahora tan excitados? Quizá sea esta expectativa lo que nos excita, pero ¿por qué pensamos que ahora es el tiempo de decidir lo que puede ser posible en principio? O estoy equivocado: ¿Es la cuestión «en principio» realmente la cuestión importante que hay que discutir ahora? Y si lo es, ¿tienen los partidarios de la IA algo importante que decirnos sobre eso?

El modelo computacional de la mente se asocia ahora con la IA, pero no es exclusivo de la IA (Noam Chomsky, por lo que yo sé, no es optimista sobre la IA, pero comparte con la IA el modelo de la computadora[1]) y el modelo de la computadora no fue inventado por la IA. Si fue inventado por alguien, lo fue por Alan Turing. La computación científica no es lo mismo que la IA.

De hecho, la idea de la mente como una especie de máquina de rememoración se remonta al siglo diecisiete.[2] A comienzos del siglo veinte dos gigantes de la lógica, Kurt Gödel y Jacques Herbrand, propusieron por primera vez la concepción moderna de la computabilidad (bajo el nombre de «recursividad general»[3]). Turing reformuló la noción de Gödel-Herbrand de la computabilidad en términos que la conectan directamente con las computadoras digitales (¡que no se habían inventado aún, sin embargo!); y también sugirió sus computadoras abstractas como modelo de la mente.[4] Aun si la

---

Hilary Putnam. Profesor Walter Beverly Pearson de matemáticas modernas y lógica matemática en el Departamento de filosofía en la Universidad de Harvard.

sugerencia de Turing probara ser errónea (aun si probara ser de alguna manera más vacía de lo que parece) todavía seguiría siendo una gran contribución al pensamiento, en la forma en que los modelos de la mente del pasado han probado ser grandes contribuciones al pensamiento; grandes intentos, aunque no sean exitosos en última instancia, de comprender la comprensión misma. Pero la IA no es la teoría de la recursión, no es la teoría de las máquinas de Turing, no es la filosofía de Alan Turing, sino que es algo mucho más específico.

Para llegar a la IA, primero tenemos que llegar a las computadoras. La moderna computadora digital es una materialización de la idea de una máquina de Turing universal en una forma particularmente efectiva, efectiva en términos de tamaño, costo, velocidad, etcétera. La construcción y el perfeccionamiento de computadoras en términos tanto de software como de hardware es un hecho de la vida. Pero no todo el que está involucrado en el diseño de software o de hardware es un investigador de IA. Sin embargo, algunas cosas por las que la IA ha obtenido crédito (por ejemplo, la inmensa mejora en las capacidades de las computadoras jugadoras de ajedrez) es tanto o más debido a los descubrimientos de los inventores de hardware que a cualquier cosa que pueda llamarse un descubrimiento de la IA.

El diseño de computadoras es una rama de la ingeniería (incluso cuando lo que se diseña es software y no hardware), y la IA es una subrama de esta rama de la ingeniería. Si vale la pena decir esto, es porque la IA se ha hecho notoria por formular reivindicaciones exageradas; reivindicaciones en el sentido de ser una disciplina fundamental e incluso de constituir «epistemología». El objetivo de esta rama de la ingeniería es desarrollar software que permita a las computadoras simular o duplicar los logros de lo que intuitivamente reconocemos como «inteligencia».

Pienso que esta es una caracterización incontrovertible de la IA. Espero que la próxima afirmación sí sea más polémica: hasta hoy la IA ha elaborado muchas cosas que son de un interés real para la computación científica en general, pero nada que arroje ninguna luz real sobre la mente (más allá de cualquier luz que hayan arrojado las discusiones de Turing). No me propongo desperdiciar mis páginas defendiendo esta última afirmación (Joseph Weizenbaum ya ha hecho un buen trabajo a lo largo de estas líneas[5]), pero daré un par de ilustraciones de lo que quiero decir.

Hace muchos años yo estaba en un simposio con uno de los

«nombres famosos» en IA. El nombre famoso había sido debidamente «modesto» sobre los logros de la IA. Dijo de repente: «Realmente no hemos logrado mucho, pero yo diría que ahora tenemos *máquinas que comprenden cuentos infantiles*». Contesté: «Conozco el programa al que usted se refiere» (era uno de los primeros programas de reconocimiento del habla). «Lo que usted no menciona es que el programa debe ser *revisado* por cada nuevo cuento infantil» (es decir, en caso de que no se haya captado el punto, el «programa» era un programa para contestar preguntas sobre un cuento infantil específico, y no un programa para comprender cuentos infantiles en general). El nombre famoso abandonó toda la cuestión en apuros.

En la actualidad, el logro más ponderado de la IA son los «sistemas expertos». Pero estos sistemas (que son, en el fondo, sólo buscadores de alta velocidad en bases de datos) no son modelos de ninguna capacidad mental interesante.

Por supuesto, sigue en pie la posibilidad de que alguna idea soñada en un laboratorio de IA pueda revolucionar en el futuro nuestros pensamientos sobre algún aspecto del uso de la mente. (El procesamiento distribuido en paralelo está excitando el interés como un modelo posible de por lo menos algunos procesos mentales, por ejemplo. Esto no es de sorprender, sin embargo, ya que el modelo fue sugerido en primer lugar por la obra del neurólogo D. O. Hebb[6]). Mi objetivo aquí no es predecir el futuro, sino sólo explicar por qué me inclino a preguntar: ¿Sobre qué es todo ese alboroto *ahora*? ¿Por qué toda una edición de *Dædalus*? ¿Por qué no esperamos que la IA logre algo y sólo entonces hacemos una edición?

«En principio»/«En la práctica»

Quizá la cuestión que interesa a la gente es si podemos modelizar la mente o el cerebro como una computadora digital (en principio tan diferentes ahora); y quizá la IA quede comprometida porque la gente no distingue con claridad la pregunta en-principio de la pregunta empírica: ¿Logrará la IA modelizar de esa manera el cerebro o la mente? Puede ser útil que comencemos viendo qué diferentes son las dos preguntas.

Por un lado la diferencia parece obvia: estamos tentados de decir que en principio puede ser posible modelizar la mente o el cerebro como una computadora digital con el software apropiado,

pero que puede ser muy difícil en la práctica escribir el software correcto. O sólo parece que esta diferencia fuese obvia. Quiero decir que pisen despacio; las cosas no son tan simples: en cierto sentido, cualquier sistema físico se puede modelizar como una computadora.[7] La afirmación de que el cerebro se puede modelizar como una computadora es por ende, en cierto sentido, trivial. Quizás haya algún otro sentido en el que podamos preguntar: ¿se puede modelar el cerebro como una computadora? En este punto, sin embargo, todo lo que podemos decir es que no está claro el sentido de esta pregunta.

Pero el sentimiento parece ser que no sólo es posible en principio modelizar la mente o el cerebro computacionalmente, sino que hay muy buenas perspectivas de que podamos hacerlo en la práctica, y los filósofos (y los desertores de la IA como Weizenbaum) se ven como reaccionarios que buscarán persuadirnos de que ni siquiera intentemos algo que promete ser un gran éxito intelectual y práctico. Si es así como uno piensa, entonces el hiato entre las dos preguntas (y la vaguedad de la pregunta en-principio) puede no ser muy importante en la práctica. Más aún, puede ser un beneficio estratégico confundirlas.

Las razones por las que podemos esperar tener éxito en la práctica no están claras para mí, sin embargo.[8] Si nosotros somos computadoras digitales programadas por la evolución, es importante entonces saber qué pensar sobre la evolución. El gran biólogo evolucionista François Jacob comparó una vez la evolución con un latonero.[9] No se debe pensar, escribió Jacob, que la evolución es como un diseñador que se sienta y produce una primorosa heliografía y luego construye organismos de acuerdo con esa heliografía. Se debe pensar más bien que la evolución es como un latonero con un depósito lleno de partes sueltas, de «basura» interesante, etc. Cada tanto el latonero tiene una idea: «Me pregunto si funcionará si pongo esta rueda de bicicleta en este torno». Muchas de las ideas brillantes del latonero fracasan, pero cada tanto alguna funciona. Los resultados son organismos con tantos rasgos arbitrarios como rasgos coherentes.

Ahora bien, imaginemos que el latonero se hace programador. Pensando todavía como un latonero, desarrolla «inteligencia natural» no escribiendo un Gran Programa y construyendo luego un dispositivo para materializarlo, sino introduciendo un dispositivo o una idea de programación después de otra. (La gente religiosa a

menudo rechaza esta concepción, porque piensa que si es correcta, toda nuestra naturaleza y nuestra historia son un «azar ciego»; pero nunca hemos podido simpatizar con esta objeción. La Providencia bien puede trabajar a través de lo que Kant llamó «la astucia de la Naturaleza».) El resultado neto sería una inteligencia que no es la expresión de algún programa, sino la expresión de billones de bits de «hojalatería».

Por cierto, alguna vez se discutió algo parecido dentro de la misma comunidad de la IA. La comunidad vacilaba entre buscar un Programa Maestro (hace diez o quince años se buscaba algo llamado lógica inductiva) y aceptar la noción de que «la inteligencia artificial es una maldita cosa después de la otra». Mi opinión es que si la IA es «una maldita cosa después de la otra», el número de «malditas cosas» en que podría pensar el latonero sería astronómico.[10] El resultado final es por cierto pesimista: si no hay Programa Maestro, luego nunca iremos demasiado lejos en términos de la simulación de la inteligencia humana. (Por supuesto, algunas áreas que han estado relativamente cerradas —por ejemplo, la demostración de teoremas en las matemáticas puras— pueden ser más dóciles. Extrañamente, la demostración de teoremas siempre ha sido una parte de la investigación en IA que ha disfrutado de subsidios financieros más bien escasos.)

¿Un programa maestro?

Pero ¿por qué no debería haber un Programa Maestro? En el caso de la lógica deductiva, hemos descubierto un conjunto de reglas que formalizan satisfactoriamente la inferencia válida. En el caso de la lógica inductiva, no hemos encontrado esas reglas, y vale la pena detenerse a preguntar por qué.

En primer lugar, no se sabe cuál se supone que es la magnitud de la lógica inductiva. Algunos escritores consideran el «método hipotético deductivo» (es decir, la inferencia que va desde el éxito de una predicción de la teoría a la aceptabilidad de la teoría) la parte más importante de la lógica inductiva, mientras otros consideran que pertenece a un asunto completamente distinto. Por supuesto, si por «inducción» queremos decir cualquier método de inferencia válido que no sea deductivo, la magnitud del tópico «lógica inductiva» sería enorme.

Si el éxito de un número grande de predicciones (digamos, mil o diez mil) que no sean ellas mismas sólo consecuencias de hipótesis auxiliares (y agregaría Karl Popper que fueran improbables en relación con lo que nos da el conocimiento de base[11]), confirmara siempre una teoría, entonces por lo menos algunas inferencias hipotético-deductivas serían fáciles de formalizar. Pero en seguida surgen problemas. Algunas teorías se aceptan cuando el número de predicciones confirmadas es aún muy pequeño. Este fue el caso con la teoría general de la relatividad, por ejemplo. Para dar cuenta de esos casos, postulamos que no sólo es el número de predicciones confirmadas lo que importa, sino también la elegancia o simplicidad de la teoría en cuestión. ¿Pueden formalizarse realmente nociones casi estéticas, como «elegancia» y «simplicidad»? Por cierto se han propuesto algunas mediciones formales, pero no puede decirse que hayan arrojado alguna luz sobre la inferencia científica en la vida real. Más aún, una teoría confirmada a veces encaja mal en nuestro conocimiento de base; en algunos casos llegamos a la conclusión de que la teoría no puede ser verdadera, mientras que en otros llegamos a la conclusión de que el conocimiento de base debe modificarse. Una vez más, fuera de parloteos imprecisos sobre la simplicidad, es difícil decir qué es lo que determina si es mejor en un caso particular preservar el conocimiento de base o modificarlo. Incluso una teoría que conduce a un vasto número de predicciones exitosas puede no aceptarse si alguien señala que una teoría mucho más simple podría también llevar a esas mismas predicciones.

A la vista de estas dificultades, algunos estudiosos de la lógica inductiva confinarían los alcances de su temática a inferencias más simples, tales como la inferencia a partir de las estadísticas de una muestra de una población a la totalidad de la población. Cuando la población consiste en objetos que existen en momentos diferentes, incluidos tiempos futuros, la muestra presente nunca llega a ser una selección al azar de la población total, sin embargo, de modo que el quid de la cuestión es: Tengo una muestra que es una selección al azar de los miembros de una población que existe ahora (o peor, de quienes existen aquí, en la Tierra, en los Estados Unidos, en el sitio en particular en el que he podido recoger las muestras, o donde sea). ¿A qué conclusión puedo llegar sobre los miembros futuros de esa población (y sobre las propiedades de los miembros en otros lugares)?

Si la muestra es una muestra de átomos de uranio, y los miembros futuros están próximos y no en un futuro cosmológico, estamos preparados para creer que los miembros futuros se parecerán a los miembros actuales, en promedio. Si la muestra es una muestra de personas, y si los miembros futuros de la población no están en un futuro próximo, es menos probable entonces partir de este supuesto, al menos si hay rasgos culturalmente variables en cuestión. Aquí nos guía el conocimiento de base, desde luego. Esta clase de ejemplos ha sugerido a algunos investigadores que quizá todo lo que hay en la inducción es el uso habilidoso del conocimiento de base: sólo nos facilita pasar de lo que sabemos a algún conocimiento adicional. Pero entonces los casos en que no tenemos el suficiente conocimiento de base, así como los casos excepcionales en los que todo lo que tenemos que hacer es precisamente interrogar ese conocimiento de base, asumen gran importancia: y aquí, como ya señalamos, nadie tiene nada que decir aparte de parloteos vagos sobre la simplicidad.

El problema de la inducción no es de ningún modo el único problema que confronta cualquiera que intente simular seriamente la inteligencia humana. La inducción —y por cierto, toda la cognición— presupone la habilidad de reconocer similitudes entre cosas; pero las similitudes de ningún modo son sólo constancias del estímulo físico o patrones en el insumo de los órganos sensoriales. Lo que hace similares a los cuchillos, por ejemplo, no es que se parezcan (pues no se parecen), sino que todos están fabricados para cortar o perforar (aquí dejo al margen casos tales como los cuchillos ceremoniales, por ejemplo). De este modo, cualquier sistema que puede reconocer los cuchillos como cosas relevantemente similares debe ser capaz de atribuir *propósito* a los agentes. Los humanos no tienen dificultades para hacerlo. Pero no está claro si lo hacemos mediante una inducción no asistida; bien podríamos tener una habilidad «precableada» para ponernos en los zapatos de otras personas que nos permite atribuirles cualquier propósito de los que seamos capaces de atribuirnos a nosotros mismos: una habilidad con que la Evolución Latonera encuentra conveniente dotarnos, una habilidad que nos ayuda a saber cuál entre las infinitamente muchas posibles inducciones podemos considerar que puede tener éxito. Una vez más, reconocer que un gran danés y un chihuahua son similares en el sentido de pertenecer a la misma especie requiere una habilidad para

darse cuenta de que, a pesar de las apariencias,[12] los chihuahuas pueden preñar a los grandes daneses y producir descendencia fértil. Pensar en términos de potencial para el apareamiento y para la reproducción es natural para nosotros, pero no necesita serlo para una inteligencia artificial, a menos que deliberadamente simulemos esta propensión humana cuando la construimos. Estos ejemplos pueden ampliarse indefinidamente.

Las similitudes que se expresan mediante adjetivos y verbos, más que mediante nombres, pueden ser todavía más complejas. Una inteligencia no humana puede saber que el color «blanco» está en una tarjeta de colores, por ejemplo, sin ser capaz de comprender por qué los humanos rosados grisáceos son llamados blancos, y puede saber qué es abrir una puerta sin ser capaz de comprender por qué hablamos de abrir una frontera o abrir relaciones comerciales. Hay muchas palabras (como lo señaló Ludwig Wittgenstein[13]) que se aplican a cosas que sólo tienen con otras un «aire de familia»; no tiene que ser una cosa que todas las *x* tienen en común. Por ejemplo, hablamos de los jefes tribales cananeos del Antiguo Testamento como de reyes, aunque sus reinos probablemente no fueran más que aldeas, y hablamos de Jorge VI como de un rey aunque él nunca llegó a reinar sobre Inglaterra; incluso decimos que, en algunos casos históricos, el reinado no fue hereditario. De la misma manera (en el ejemplo de Wittgenstein), no hay ninguna propiedad que todos los juegos tengan en común que los distingan a todos de las cosas que no son juegos.

El trabajo categorial de la inteligencia artificial es simular la inteligencia, no duplicarla. De modo que quizá se podrían afinar los problemas mencionados construyendo un sistema que razone en un lenguaje ideal[14], un lenguaje en el que las palabras no cambiaran sus extensiones dependiendo del contexto (en ese lenguaje una hoja de papel mecanográfico puede ser «blanco_1» y los humanos «blanco_2», donde «blanco_1» es el color de la tarjeta de colores y «blanco_2» es rosa grisáceo). Quizá todas las palabras con aire de familia deban ser barridas de ese lenguaje (¿cuánto vocabulario quedaría?). Pero mi lista de dificultades aún no termina.

Dado que el proyecto de la lógica simbólica inductiva pareció salirse de madre después de Rudolf Carnap, el pensamiento entre los filósofos de la ciencia , como lo he reportado, corrió en dirección de los métodos que atribuyen una gran importancia al conocimiento de base. Es instructivo ver por qué los filósofos han adoptado esta

estrategia y también darse cuenta de qué insatisfactoria es, si nuestro propósito es simular la inteligencia en lugar de describirla.

Un problema pesado se puede describir como la existencia de inducciones conflictivas. Aquí hay un ejemplo de Nelson Goodman: en la medida en que yo sé, nadie que haya ingresado al Emerson Hall en la Universidad de Harvard ha sido capaz de hablar la lengua inuit (esquimal). Esta afirmación sugiere la inducción de que si alguien entra al Emerson Hall, luego él o ella no habla inuit.[15] Pongamos que Ukuk sea un esquimal de Alaska que habla inuit. ¿Debo predecir que si Ukuk entra en el Emerson Hall, Ukuk no podrá hablar inuit nunca más? Obviamente no, pero ¿qué hay de erróneo en esta inducción?

Goodman responde que lo que anda mal con esta inferencia es que se halla en conflicto con la ley inductivamente sustentada y «mejor engranada», que dice que la gente no pierde su habilidad de hablar una lengua porque entre a un nuevo lugar. Pero ¿cómo puedo saber yo que esta ley tiene más instancias confirmatorias que la regularidad que dice que nadie que entra en el Emerson Hall habla inuit? ¿De nuevo a través del conocimiento de base?

De hecho, no creo que cuando yo era chico tuviera alguna idea de la frecuencia con que se hubieran confirmado las regularidades en conflicto de ese ejemplo (conflictivas, en la medida en que una de ellas ha de fallar cuando Ukuk entre en el Emerson Hall); pero aun así yo sabía lo suficiente como para no hacer la tonta inducción de que Ukuk dejaría de hablar inuit si entraba en un edificio (o en un país) donde nadie supiera hablar inuit. Una vez más, no está claro que el conocimiento de que uno no pierde una lengua sea realmente producto de una inducción; quizá sea algo que tenemos una propensión innata a creer. La cuestión que no desaparecerá es *cuánto de lo que llamamos inteligencia presupone al resto de la naturaleza humana.*

Además, si lo que realmente importa es el «engranamiento» (es decir, el número y la variedad de instancias confirmadoras), y si la información de que la proposición universal «Uno no pierde la capacidad de hablar una lengua cuando entra en un lugar» está mejor engranada que la proposición universal «Ninguno que entra en el Emerson Hall habla inuit» es parte de mi conocimiento de base, no está para nada claro cómo es que esa información llegó allí. Quizá la información esté implícita en la forma en que la gente habla de las capacidades lingüísticas; pero entonces uno se

enfrenta con el problema de cómo decodifica la información implícita que transmiten las elocuciones que se escuchan.

El problema de las inducciones conflictivas es ubicuo, incluso si se restringe la atención a las inferencias inductivas más simples. Si la solución consistiera sólo en darle al sistema más conocimiento de base, ¿cuáles son entonces sus implicaciones para la inteligencia artificial?

No es fácil de decir, porque la inteligencia artificial como la conocemos realmente no trata en absoluto de simular inteligencia. La simulación de la inteligencia es sólo su actividad categorial; su actividad real consiste en escribir programas sagaces para una variedad de tareas. Pero si la inteligencia artificial existiera como una actividad de investigación real, y no categorial, habría entonces dos estrategias alternativas que podrían adoptar sus practicantes cuando se enfrentaran con el problema del conocimiento de base:

1. Podrían aceptar la perspectiva de los filósofos de la ciencia que he descrito, tratando simplemente de programar en la máquina toda la información que posee un humano inductivo sofisticado (incluyendo la información implícita). Se requerirían por lo menos varias generaciones de investigadores para formalizar la información (probablemente no pueda hacerse en absoluto, debido a la cantidad de información involucrada), y no está claro que el resultado vaya a ser algo más que un gigantesco sistema experto. Nadie encontraría esto demasiado excitante, y una «inteligencia» semejante sería con toda probabilidad terriblemente poco imaginativa, incapaz de darse cuenta de que en muchos casos es precisamente conocimiento de base lo que se le debe dar.

2. Los practicantes de la IA podrían afrontar la tarea mucho más excitante y ambiciosa de construir un dispositivo que pueda aprender el conocimiento de base mediante la interacción con seres humanos, lo mismo que un niño aprende una lengua y toda la información cultural, explícita e implícita, que viene con el aprendizaje de un lenguaje cuando se crece en una comunidad humana.

El problema del lenguaje natural

La segunda alternativa es por cierto el proyecto que merece el nombre de «inteligencia artificial». Pero consideremos los proble-

mas: para darse cuenta de cuál es la información implícita en las cosas que la gente dice, la máquina debe simular la comprensión de un lenguaje humano. De este modo, si se adopta esta estrategia, hay que abandonar la idea de pegarse a un lenguaje ideal artificial y de ignorar las complejidades del lenguaje natural, y hay que abandonarla porque el costo es muy alto. Una enorme parte de la información que necesitaría la máquina sólo puede recuperarse a través del procesamiento en lenguaje natural.

Pero el problema del lenguaje natural presenta de nuevo muchas dificultades. Chomsky y su escuela creen que hay un «molde» innato para el lenguaje natural, incluyendo los aspectos «semánticos» o conceptuales, y que está precableado por Evolución, el Latonero.[16] Aunque esta concepción es llevada al extremo por Jerry Fodor, quien sostiene que existe un lenguaje del pensamiento innato, con primitivas adecuadas para la expresión de todos los conceptos que los humanos son capaces de aprender a expresar en un lenguaje natural,[17] Chomsky mismo ha dudado en ir tan lejos. Chomsky parece comprometido con la existencia de un gran número de habilidades conceptuales innatas que nos dan una propensión a formar ciertos conceptos y no otros. (En una conversación, él ha sugerido que la diferencia entre postular conceptos innatos y postular capacidades innatas no es importante si las habilidades postuladas son lo suficientemente estructuradas.) En el extremo opuesto se halla la concepción del conductismo clásico, que explica el aprendizaje lingüístico como un caso especial de la aplicación de reglas generales para adquirir «hábitos», es decir, como un paquete más de inducciones. (Es posible, por supuesto, una postura intermedia: ¿por qué el aprendizaje del lenguaje no puede depender en parte de heurísticas de propósito especial y en parte de estrategias de aprendizaje generales, ambas desarrolladas a través de la evolución?.)

Consideremos la concepción que dice que el aprendizaje lingüístico no es realmente aprendizaje, sino más bien la maduración de una capacidad innata en un ambiente específico (algo así como la adquisición de un canto de llamada por parte de un pájaro joven que debe oír la llamada por parte de adultos de la especie, pero que también posee una propensión innata a adquirir esa clase de llamada). En su forma extrema, esta concepción conduce al pesimismo sobre la probabilidad de que se pueda simular con éxito en una computadora el uso humano del lenguaje natural.

Esta es la razón por la cual Chomsky es pesimista a propósito de los proyectos del procesamiento computacional del lenguaje natural, aunque comparte el modelo computacional de la mente, o al menos el «órgano del lenguaje» con los investigadores en IA. Nótese que esta concepción pesimista sobre el aprendizaje lingüístico es paralela a la concepción pesimista que sostiene que la inducción no es una habilidad singular, sino más bien la manifestación de una compleja naturaleza humana cuya simulación computacional requeriría un vasto número de subrutinas, tan vasto que se requerirían generaciones de investigadores para formalizar aunque más no sea una pequeña parte de ese sistema.

De la misma manera, la concepción optimista que afirma que hay un algoritmo de tamaño manejable para la lógica inductiva es paralela a la concepción optimista respecto del aprendizaje lingüístico. Esta es la idea de que hay una heurística para el aprendizaje más o menos neutral con respecto al tópico, y que esta heurística alcanza (sin la ayuda de un repertorio inmanejablemente grande de conocimiento de base precableado o de habilidades conceptuales específicas con respecto al tópico) para aprender el lenguaje natural propio, así como para hacer inferencias inductivas. Quizá la concepción optimista tenga razón, pero yo no veo ninguna en escena, sea en la inteligencia artificial o en la lógica inductiva, que posea ideas interesantes sobre la forma en que trabaja la estrategia de aprendizaje neutral con respecto al tópico. Cuando alguien aparezca con alguna de esas ideas, ésa será la ocasión para que *Dædalus* publique una edición sobre IA.

Notas

[1] Noam Chomsky, *Modular Approaches to the Study of the Mind* (San Diego, Calif.: San Diego State University Press, 1983).

[2] Esto se describe muy bien en Justin Webb, *Mechanism, Mentalism, and Metamathematics* (Dordrecht: Reidel, 1980).

[3] La concepción de recursividad de Gödel-Herbrand fue desarrollada luego por Stephen Kleene, Alonzo Church, Emil Post y Alan Turing. La identificación de la recursividad con la computabilidad efectiva fue sugerida (aunque en forma oblicua) por Kurt Gödel en «Sobre sentencias formalmente indecidibles en *Principia Mathematica* y sistemas afines». El original alemán de este artículo se publicó en el *Monatshefte für Mathematik und Physik* 38 (1931): 173-98; la traducción inglesa está en *The Undecidable: Basic Papers on Undecidable Propositions, Undecidable*

Problems, and Computable Functions, ed. de Martin Davis (Hewlett, N.Y.: Raven Press, 1965), 5-38. La idea fue explícitamente adelantada por Church en su ensayo clásico sobre la indecidibilidad de la aritmética, «A Note on the Entscheidungsproblem», *Journal of Symbolic Logic* 1 (1) (marzo de 1936):40-41; corrección, ibid. (3) (setiembre de 1936):101-102; reimpreso en Davis, *The Undecidable*, 110-15.

[4] Alan Turing y Michael Woodger, *The Automatic Computing Machine: Papers by Alan Turing and Michael* Woodger (Cambridge: MIT Press 1985).

[5] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: Freeman, 1976).

[6] Véase David E. Rumelhart, James L. McClelland y el PDP Research Group, editores, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vols. 1 y 2 (Cambridge: MIT Press, 1986); y D. O. Hebb, *Essay on Mind* (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980).

[7] Más precisamente, si estamos interesados en la conducta de un sistema físico que sea finito en el espacio y en el tiempo y si deseamos predecir esa conducta sólo hasta cierto nivel especificado de exactitud, entonces (suponiendo que las leyes del movimiento son funciones continuas) es trivial mostrar que una función discreta dará una predicción con el nivel de exactitud especificado. Si los valores posibles de los parámetros de límite se restringen a un rango finito, entonces un conjunto finito de tales funciones discretas dará la conducta del sistema bajo todas las condiciones posibles en el rango especificado dentro de la exactitud deseada. Pero si éste fuera el caso, la conducta del sistema se describe mediante una función recursiva y de allí que el sistema pueda ser modelizado mediante un autómata.

[8] En su respuesta a este artículo (en esta misma edición), Daniel Dennett me acusa de ofrecer una argumentación «a priori» de que el éxito es imposible. No he cargado en absoluto el texto de este ensayo a la luz de esta respuesta e invito al lector a observar que yo no postulo ninguna afirmación de ninguna «prueba a priori de imposibilidad». Aunque Dennett dice que él va a explicar qué nos ha enseñado la IA sobre la mente, lo que de hecho hace es repetir los insultos que los investigadores de IA arrojan a los filósofos («¡Nosotros somos los experimentadores, ustedes los pensadores de sillón!»). En otras ocasiones, cuando Dennett no habla como portavoz de la IA sino que hace lo que hace mejor, que es filosofía, él está bien al tanto, por supuesto, de que yo y, para el caso, otros filósofos que él respeta, no estamos para nada enredados en un razonamiento a priori, y que el hecho de que no hacemos «experimentos» no significa que no estemos comprometidos (como él lo está) en pensar sobre el mundo real a la luz del mejor conocimiento disponible.

[9] François Jacob, «Evolution and Tinkering», *Science* 196 (1977):1161-66.

[10] Que el número de veces que nuestro diseño haya sido modificado por la evolución pueda ser astronómico ni significa que las modificaciones exitosas no estén (parcialmente) organizadas en forma jerárquica, ni significa que no haya muchos principios que expliquen el funcionamiento global de todos los componentes. Describir la alternativa al éxito de la IA como «la mente del caos», como lo hace Dennett, es un sinsentido. Si resultara ser que nuestra mente es un caos cuando se la modeliza como computadora, eso sólo mostraría que el formalismo computacional no es un formalismo perspicuo para describir el cerebro, y no que el cerebro sea un caos.

[11] Karl Popper, *The Logic of Scientific Discovery* (Londres: Hutchinson, 1959).

[12] Nótese que, si sólo dispusiéramos de las apariencias para orientarnos, sería muy natural considerar a los grandes daneses y a los chihuahuas como animales de diferentes especies.

[13] Véase Ludwig Wittgenstein, *Philosophical Investigations* (Oxford: Basil Blackwell, 1958), sec. 66-71.

[14] Nótese que esta idea era una de las piedras fundamentales del positivismo lógico. Aunque el objetivo de los positivistas era reconstruir el razonamiento científico más que mecanizarlo, se precipitaron en todos y cada uno de los problemas aquí mencionados; en muchos aspectos, la historia de la inteligencia artificial es una repetición de la historia del positivismo lógico (la segunda vez quizá como farsa).

[15] Nelson Goodman, *Fact, Fiction, and Forecast*, 4ª ed. (Cambridge: Harvard University Press, 1983).

[16] Chomsky habla de «un subsistema [para el lenguaje] que posee un carácter integrado específico y que es en efecto el programa genético para un órgano específico» en la discusión con Seymour Papert, Jean Piaget y otros, reimpreso en *Language and Learning*, ed. de Massimo Piatelli (Cambridge: Harvard University Press, 1980). Véase también Noam Chomsky, *Language and Problems of Knowledge, The Managua Lectures* (Cambridge: MIT Press, 1987).

[17] Jerry A. Fodor, *The Language of Thought* (Nueva York: Thomas Y. Crowell, 1975).

13
Cuando los filósofos se encuentran con la inteligencia artificial

Daniel C. Dennett

¿Cómo es posible que una forma física —una persona, un animal, un robot— extraiga conocimiento del mundo a partir de la percepción y luego explote ese conocimiento en la guía de acciones exitosas? Esta es una pregunta con la que los filósofos han lidiado durante generaciones, pero también se puede considerar una de las preguntas definitorias de la inteligencia artificial. La IA es, en gran medida, filosofía. A menudo está directamente involucrada en preguntas filosóficas inmediatamente reconocibles: ¿Qué es la mente? ¿Qué es el significado? ¿Qué son el razonamiento y la racionalidad? ¿Cuáles son las condiciones necesarias para el reconocimiento de objetos en la percepción? ¿Cómo se toman y cómo se justifican las decisiones?

Algunos filósofos han apreciado este aspecto de la IA, y unos pocos han cambiado gustosos de campo para luchar esas querellas filosóficas a través de matorrales de LISP. (El lenguaje de programación LISP, creado por John McCarthy, es la lingua franca de la IA.) En general, sin embargo, los filósofos no han dado la bienvenida a este nuevo estilo de filosofía con demasiado entusiasmo. Uno puede suponer que es porque han visto a través de ella. Algunos filósofos por cierto han llegado a la conclusión, luego de una inspección sumaria del campo, que a despecho de las pasmosas pretensiones de algunos de sus publicistas, la inteligencia artificial no tiene nada nuevo que ofrecer a los filósofos aparte del espectáculo de viejos y bien fatigados errores, cometi-

Daniel C. Dennett. Profesor emérito de artes y ciencias y profesor de filosofía en la Universidad de Tufts. Es director del Centro de Estudios Cognitivos y codirector del Estudio de Software Curricular, ambos en la misma universidad.

dos de nuevo en un medio nuevo. Y otros filósofos están tan seguros de que así debe ser que ni siquiera se han molestado en conducir una inspección sumaria. Están seguros de que se puede desechar el campo sobre la base de «principios generales».

Los filósofos han soñado con la IA durante siglos. Hobbes y Leibniz, en formas muy diferentes, trataron de explorar las implicaciones de la idea de particionar la mente en operaciones pequeñas y en última instancia mecánicas. Descartes anticipó incluso la prueba de Turing (la tan discutida propuesta de Alan Turing de una audiencia de forma para la computadora, en la que el trabajo de la computadora es convencer a los jueces de que están conversando con un ser humano[1]) y no dudó en formular una confiada predicción de su inevitable resultado:

> Es por cierto concebible que se pueda hacer una máquina de modo tal que pueda proferir palabras, e incluso palabras apropiadas a la presencia de actos u objetos físicos que causen algún cambio en sus órganos; como, por ejemplo, si fuera tocada en alguna parte, que preguntara qué se está intentando decirle; si se la tocara en otra, que gritara que ha sido herida, y así para cosas similares. Pero nunca podría modificar sus frases para responder al sentido de lo que se dijera en su presencia, como incluso el más estúpido de los hombres puede hacer.[2]

La apreciación que tenía Descartes de los poderes del mecanismo se hallaba teñida por su conocimiento de los maravillosos autómatas de relojería de su época. Podía ver muy clara y distintamente, sin duda, las limitaciones de esa tecnología. ¡Ni aun mil pequeños engranajes —ni aun diez mil— permitirían al autómata responder completa y racionalmente! Quizás Hobbes y Leibniz hubieran sido menos confiados en ese punto, pero seguramente ninguno de ellos se habría molestado en preguntarse sobre las limitaciones a priori de un millón de pequeños engranajes girando millones de veces por segundo. Para ellos sencillamente ese no era un pensamiento pensable. Era impensable, entonces, no en el sentido filosófico familiar de parecer autocontradictorio («repugnante a la razón») o enteramente fuera de su esquema conceptual (como el concepto de neutrino), sino en el sentido más cotidiano, pero por igual limitante, de ser una idea que no había forma de tomar en serio. Cuando los filósofos se asoman a escudriñar grandes dominios conceptuales, se encuentran tan inhibidos en los caminos de su sentido de la tontería

como por su intuición de la necesidad lógica. Y hay algo a propósito de la IA que muchos filósofos encuentran despreciable; si no repugnante a la razón, repugnante a su sentido estético.

Este choque de visiones se exhibió en forma memorable en un debate histórico en la Universidad de Tufts en marzo de 1978, escenificado, apropiadamente, por la Sociedad para la Filosofía y la Psicología. Nominalmente un panel de discusión sobre las fundamentaciones y perspectivas de la inteligencia artificial se convirtió en una lucha de equipo entre cuatro ideólogos de peso pesado: Noam Chomsky y Jerry Fodor atacando a la IA, y Roger Schank y Terry Winograd defendiéndola. En esa época Schank estaba trabajando en programas para la comprensión del lenguaje natural, y los críticos concentraron en su esquema para la representación (en una computadora) la confusa colección de trivialidades que todos conocemos y que de alguna manera descansan en la forma en que desciframos actos de habla ordinarios, alusivos y truncados como lo son. Chomsky y Fodor amontonaron burlas contra esa empresa, pero las bases de su ataque cambiaron gradualmente en el curso del enfrentamiento. Comenzó como una condena directa del error, basada en los «primeros principios» —Schank se hallaba en un error tonto o en otro— pero finalizó con una sorprendente concesión por parte de Chomsky: podría llegar a ser, como Schank pensaba, que la capacidad humana para comprender la conversación (y más generalmente, para pensar) debiera explicarse en términos de la interacción de cientos o miles de expresiones edificadas a la diabla (seudorrepresentaciones, se las podría llamar), pero eso sería una vergüenza, porque entonces la psicología demostraría en último análisis no ser «interesante». Sólo podía haber dos posibilidades interesantes, a entender de Chomsky: la psicología podía volverse «como la física» (sus regularidades se explicarían como las consecuencias de unas pocas leyes profundas, elegantes e inexorables) o la psicología podía volverse absolutamente carente de leyes, en cuyo caso la única forma de estudiar o exponer la psicología sería a la manera de los novelistas (y él prefería ampliamente a Jane Austen que a Roger Schank, si de eso se trataba).

Siguió un vigoroso debate entre los panelistas y los asistentes, coronado por una observación del colega de Chomsky en el Massachusetts Institute of Technology, Marvin Minsky, uno de los padres fundadores de la IA y fundador del Laboratorio de Inteligencia Artificial del MIT: «Pienso que sólo un profesor de humani-

dades del MIT podría ser tan negligente a la tercera posibilidad interesante: que la psicología se vuelva como la ingeniería».

Minsky había puesto el dedo en la llaga. Hay algo en la perspectiva de una estrategia ingenieril hacia la mente que es profundamente repugnante para cierta clase de humanismo, y eso tiene poco o nada que ver con un disgusto por el materialismo o por la ciencia. Véase si no el culto de Chomsky por la física, una actitud que comparte con muchos filósofos. Los días del idealismo a la Berkeley y del dualismo cartesiano han pasado (si se puede juzgar por el actual consenso materialista entre los filósofos y científicos), pero en su lugar hay una amplia aceptación de lo que podríamos llamar la horquilla de Chomsky: hay sólo dos («interesantes») alternativas.

Por un lado están la dignidad y la pureza de la Mente Cristalina. Recuérdese el prejuicio de Aristóteles contra la extensión a los cielos de la física terrena, la que debía, según pensaba, estar ligada a un orden más alto y más puro. Este fue su único legado pernicioso, pero ahora que los cielos han sido atormentados, apreciamos la belleza de la física universal y podemos esperar que la mente estará entre las «cosas naturales» selectas, y no que sea una mera tergiversación de pedazos y piezas.

Por el otro lado está la dignidad del último misterio, la Mente Inexplicable. Si nuestras mentes no pueden ser fundamentales, hagamos que sean anómalas. Una perspectiva muy influyente entre los filósofos en tiempos recientes ha sido el «monismo anómalo» de Donald Davidson, la perspectiva de que a pesar de que la mente es el cerebro, no hay regularidades legaliformes que pongan en línea los hechos mentales con los hechos físicos.[3] John Searle, colega de Davidson en Berkeley, ha hecho de la mente una clase diferente de misterio: el cerebro, gracias a algún rasgo no especificado de su bioquímica, posee algunos «poderes causales de abajo hacia arriba» terriblemente importantes —pero no especificados— que son completamente distintos de los meros «poderes de control» que se estudian en IA.

Un rasgo compartido por estas de otro modo drásticamente diferentes formas de materialismo de la mente-cuerpo es su resistencia frente al *tertium quid* de Minsky: entre la mente como cristal y la mente como caos está la mente como aparato, un objeto del que no se debe esperar que esté gobernado por «profundas» leyes matemáticas, pero así y todo un objeto *diseñado*, analizable en términos funcionales: fines y medios, costos y beneficios,

soluciones elegantes por un lado, y en la otra atajos, súplicas al jurado y parches baratos *ad hoc*.

Esta visión de la mente es resistida por muchos filósofos a pesar de ser una implicación directa de la concepción vigente entre los científicos y los humanistas de mente científica sobre nuestro puesto en la naturaleza: somos entidades biológicas diseñadas por la selección natural, que es un latonero y no un ingeniero ideal. Los programadores de computadoras llaman *kludges* a los parches *ad hoc* (lo que rima con *scrooge*, «avariento»), y la mezcla de desdén y envidiosa admiración que se reserva a los *kludges* es paralela a la fascinación de los biólogos por «el pulgar del panda» y otros fascinantes ejemplos de *bricolage*, para usar la palabra de François Jacob.[4] El retruécano inadvertido más perfecto que he escuchado fue proferido por Barbara Partee en una crítica caliente a un *kludge* reconocido en un analizador de lenguaje natural de IA: «¡Ese odioso parche!» La naturaleza está llena de parches odiosos, muchos de ellos perversamente brillantes. Aunque este hecho es ampliamente apreciado, sus implicaciones para el estudio de la mente son a menudo repugnantes para los filósofos, pues sus métodos apriorísticos tradicionales para investigar la mente les dan poco poder para estudiar fenómenos que pueden estar plagados de horribles parches. Realmente hay una sola forma de estudiar esas posibilidades: con el esquema mental más empírico de «ingeniería inversa».

Esta resistencia es claramente manifiesta en el ensayo de Hilary Putnam en esta edición de *Dædalus*, que puede servir como caso conveniente (si no particularmente florido) del síndrome que deseo discutir. La horquilla de Chomsky, la mente como cristal o como caos, es transformada por Putnam en un movimiento pendular que él cree observar dentro de la misma IA. El afirma que la IA se «balanceó» a través de los años entre la búsqueda del Programa Maestro y la aceptación de la idea de que «la inteligencia artificial es una maldita cosa después de la otra». Yo mismo no he observado ese balanceo en el campo a través de los años, pero pienso que sé dónde quiere llegar. Aquí tenemos, entonces, una perspectiva diferente sobre la misma cuestión.

Entre las muchas divisiones de opinión dentro de la IA, hay una facción (a veces llamada los logicistas) cuyas aspiraciones me sugieren que son los buscadores del Programa Maestro de Putnam. Han sido mejor caricaturizados recientemente por un inves-

tigador de IA como los buscadores de «las ecuaciones de Maxwell para el pensamiento». Bajo esta rúbrica se pueden amontonar muchas otras empresas del campo más o menos incompatibles. A grandes rasgos, lo que tienen en común es la idea de que debe haber no un Programa Maestro, sino que debe haber algo más parecido a un lenguaje maestro de programación, un sistema lógicamente sólido de representación explícita para todo el conocimiento que reside en un agente (natural o artificial). Pegado a esta librería de hechos representados (que, en efecto, se pueden tratar como axiomas) y operando computacionalmente sobre ella hay una clase u otra de «máquina de inferencia», capaz de deducir las implicaciones relevantes de los axiomas relevantes y eventualmente capaz de vomitar mediante este proceso de inferencia los imperativos o decisiones que inmediatamente se implementen.

Por ejemplo, supongamos que la percepción produzca la nueva premisa urgente (expresada en el lenguaje maestro de programación) que se está aproximando el borde de un precipicio; esto haría que la máquina de inferencia obtuviera de la memoria los hechos almacenados pertinentes sobre acantilados, gravedad, aceleración, impacto, daño, la suprema indeseabilidad de esos daños y los efectos probables de clavar los frenos o de seguir avanzando. Inmediatamente, se espera, la máquina de inferencia deducirá un teorema al efecto de que se haga parar, y sin más vueltas se detendrá.

La parte difícil es el diseño de un sistema de esta clase que realmente trabaje bien en tiempo real, permitiendo que se realicen millones de operaciones por segundo en la máquina de inferencia. Como el mundo reconoce que es un problema de destreza en tiempo real, lo que pone a los lógicos aparte es su convicción respecto de que la forma de resolverlo es encontrar un vocabulario verdaderamente perspicuo y una forma lógica para el lenguaje maestro. La lógica moderna ha demostrado ser un instrumento poderoso para explorar y representar el venerable universo de las matemáticas; la esperanza nada irrazonable de los lógicos es que se puede dotar al mismo sistema de lógica para que capture el universo caótico de los agentes que andan su camino en el proteico mundo macroscópico. Si usted tiene los axiomas correctos, creen ellos, el resto será fácil. El problema que ellos encuentran tiene que ver con conservar bajo el número de axiomas por razones de generalidad (que es una imposición), mientras que no se exige al sistema perder el tiempo volviendo a deducir hechos cruciales de nivel intermedio cada vez que ve un acantilado.

Esta idea de axiomatizar la realidad cotidiana es con seguridad una idea filosófica. Spinoza la habría amado, y muchos filósofos contemporáneos que trabajan en lógica filosófica y en la semántica del lenguaje natural comparten por lo menos el objetivo de delinear un sistema lógico riguroso en el que cada proposición, cada pensamiento, cada presentimiento y duda se pueda expresar inequívocamente. La idea no fue reinventada por la IA; fue un don de los filósofos que crearon la lógica matemática moderna: George Boole, Gottlob Frege, Alfred North Whitehead, Bertrand Russell, Alfred Tarski y Alonzo Church. Douglas Hofstadter llama a este tema de la IA el sueño booleano.[5] Siempre tuvo sus adherentes y sus críticos, con muchas variaciones.

El relato de Putnam sobre este tema como la búsqueda del Programa Maestro es bien claro, pero cuando él describe el polo opuesto, pasa por alto nuestras dos perspectivas restantes: la mente como aparato y la mente como caos. Como él lo expone, «Si la IA es 'una maldita cosa después de la otra', el número de 'malditas cosas' en que tiene que pensar el latonero sería astronómico. El resultado final es por cierto pesimista: si no hay Programa Maestro, luego nunca iremos demasiado lejos en términos de la simulación de la inteligencia humana». Aquí Putnam eleva una posibilidad del peor caso (el aparato será total y astronómicamente *ad hoc*) como la única alternativa probable al Programa Maestro. ¿Por qué lo hace? ¿Qué es lo que tiene en contra de la posibilidad de explorar el vasto espacio de posibilidades ingenieriles entre el Cristal y el Caos? La sabiduría biológica, lejos de favorecer este pesimismo, sustenta nuestra esperanza de que la mezcla de elegancia y Rube Goldberg que se halla en todas partes en la naturaleza (en la bioquímica de la reproducción, por ejemplo) será discernible también en la mente.

Hay, en efecto, una variedad de estrategias muy diferentes que están realizando en IA aquellos que esperan que la mente demostrará ser una especie de aparato o una colección de aparatos parcialmente integrados. Todos ellos favorecen la lógica, la austeridad y el orden en algunos aspectos de sus sistemas, y sin embargo explotan la utilidad peculiar del desenfreno, la inconsistencia y el desorden en otros aspectos. No es que los dos temas de Putnam no existan en IA, sino que, al describirlos como las únicas alternativas, impone en el campo una taxonomía de Procusto que hace difícil discernir las cuestiones interesantes que realmente lo orientan.

La mayoría de los proyectos de IA son exploraciones de *formas en que se deben hacer las cosas* y como tales son más como experimentos pensantes que como experimentos empíricos. Difieren de los experimentos pensantes de la filosofía no primariamente en sus contenidos, sino en su metodología: reemplazan algunos (no todos) los supuestos de base «intuitivos», «plausibles» y fluctuantes de los experimentos pensantes filosóficos por restricciones dictadas por la demanda de que el modelo debe correr en una computadora. Esta restricciones de tiempo y espacio y las exigencias de especificación se pueden intercambiar mutuamente en formas casi ilimitadas, de modo que se impongan nuevas «máquinas virtuales» o «arquitecturas virtuales» sobre la arquitectura serial subyacente de la computadora digital. Algunas elecciones de intercambio están mejor motivadas, son más realistas o más plausibles que otras, por supuesto, pero en todos los casos las restricciones impuestas sirven para disciplinar la imaginación (y por lo tanto las pretensiones) del experimentador pensante. Hay pocas chances de que un filósofo sea sorprendido (o más exactamente, defraudado) por los resultados de su propio experimento pensante, pero esto pasa todo el tiempo en IA.

Un filósofo que mire de cerca estos proyectos encontrará abundante base para su escepticismo. Muchos parecen basarse en esperanzas desoladas o en entusiasmos bastardos hacia rasgos arquitectónicos o de manejo de la información, y si ex-trapolamos a partir de la breve historia del campo, podemos asegurar que la mayor parte del escepticismo será vindicada tarde o temprano. Lo que convierte a la IA en una superación de los tempranos esfuerzos filosóficos en el diseño de modelos, sin embargo, es la forma en que el escepticismo se reivindica: mediante el fracaso concreto del sistema en cuestión. Igual que los filósofos, los investigadores de IA celebran toda nueva propuesta con juicios intuitivos sobre sus perspectivas, respaldados por argumentos más o menos a priori sobre las razones por las cuales cierto rasgo no puede estar ahí o no se podrá hacer que funcione. Pero, a diferencia de los filósofos, estos investigadores no quedan contentos con sus argumentos e intuiciones; reservan cierto espacio para ser sorprendidos por los resultados, una sorpresa que sólo puede ser provocada por la potencia inesperada que demuestran los sistemas maquinados concretos en acción.

Putnam revisa una panoplia de problemas que enfrenta la IA: los problemas de la inducción, de discernir las similitudes rele-

vantes, del aprendizaje, de la modelización del conocimiento de base. Son problemas ampliamente reconocidos en IA, y las puntualizaciones que hace a su respecto ya fueron hechas antes por la gente que trabaja en IA, quienes han seguido tratando de enfrentar los problemas mediante diversas propuestas relativamente concretas. Las diabólicas dificultades que Putnam observa en reseñas tradicionales del proceso de inducción, por ejemplo, han sido catalogadas en forma más incisiva por John Holland, Keith Holyoak, Richard Nisbett y Paul Thagard en su reciente libro *Induction*,[6] pero su diagnóstico de esas enfermedades es el preámbulo para el esbozo de modelos de IA diseñados para superarlos. Pueden encontrarse en abundancia modelos que afrontan los problemas de la distinción de similitudes y de los mecanismos de aprendizaje. El proyecto SOAR de John Laird, Allen Newell y Paul Rosenbloom[7] es un ejemplo estimable. Y el tema de la importancia —y dificultad— de la modelización del conocimiento de base ha sido ubicuo en los años recientes, junto con muchas sugerencias para soluciones bajo investigación. Ahora bien, quizá todos sean desesperanzados, como Putnam se inclina a creer, pero eso simplemente no puede decirse sin construir concretamente los modelos y poniéndolos a prueba.

La última afirmación no es estrictamente verdad, por supuesto. Cuando una refutación a priori de una idea está fundada, el constructor de modelos empíricos que persista a pesar de la refutación tarde o temprano tendrá que enfrentar un coro que le grita «¡Te lo dijimos!» Este es uno de los riesgos ocupacionales de la IA. El meollo de la cuestión es cómo distinguir las pruebas genuinas de imposibilidad a priori de meras fallas de la imaginación. La respuesta tradicional de los filósofos es *Más análisis y argumentación a priori.* La respuesta de los investigadores de IA es, *Construyámoslo y se verá.*

Putnam ofrece una sorprendente instancia de esta diferencia en su revisión de las posibilidades para atacar el problema del conocimiento de base. Como Descartes, se las ingenia para imaginar una ficción de experimento pensante que se está convirtiendo en real, e igual que Descartes está preparado para desecharla por anticipado. Se puede, dice Putnam,

tratar simplemente de programar en la máquina toda la información que posee un humano inductivo sofisticado (incluyendo la información

implícita). Se requerirían por lo menos varias generaciones de investigadores para formalizar la información (probablemente no pueda hacerse en absoluto, debido a la cantidad de información involucrada), y no está claro que el resultado vaya a ser algo más que un gigantesco sistema experto. Nadie encontraría esto demasiado excitante, y una «inteligencia» semejante sería con toda probabilidad terriblemente poco imaginativa...

Esto describe casi perfectamente el enorme proyecto CYC de Douglas Lenat.[8] Puede decirse que Lenat está intentando crear la proverbial enciclopedia ambulante: ¡una memoria de conocimiento de sentido común en forma de una base de datos que contiene todos los hechos expresados (o tácitamente supuestos) en una enciclopedia! Esto involucra fabricar a mano millones de representaciones en un lenguaje individual (que con el tiempo debera unificarse, un trabajo nada pequeño), a partir de las cuales se espera que la máquina de inferencia sea capaz de deducir cualquier cosa que necesita a medida que encuentra novedades en su mundo: por ejemplo, el hecho de que la gente en general prefiera no cortarse los pies, o que los baños de sol sean raros en Cape Cod en el mes de febrero.

Casi todos los líderes de opinión en IA comparten la visión prejuiciada de Putnam sobre este proyecto: no está claro, como dice Putnam, que el proyecto vaya a hacer cosas que nos enseñen algo sobre la mente; con toda probabilidad, como él dice, será terriblemente poco imaginativo. Y muchos irían más lejos e insistirán en que sus perspectivas son tan desesperanzadas y su costo tan grande que debería abandonarse en beneficio de avenidas más promisorias. (La estimación concreta se mide en *siglos*-persona de trabajo, una cifra que Putnam no se ha molestado a imaginar en detalle.) Pero el proyecto está financiado, y ya se verá.

Lo que tenemos aquí es un choque de supuestos metodológicos bien fundamentales. Los filósofos se inclinan a ver los proyectos de IA con el desprecio condescendiente que uno reserva para esos tontos persistentes que tratan de cuadrar el círculo o de trisectar el ángulo con compás y escuadra: ¡hemos *probado* que eso no puede hacerse, de modo que bótalo! Pero las pruebas no son geométricas; están enganchadas con supuestos sobre condiciones de límite «plausibles» y repletas con idealizaciones que pueden mostrarse tan irrelevantes aquí como las notorias

pruebas de los expertos en aerodinámica respecto de que los abejorros no pueden volar.

Pero uno podría seguir preguntándose, haciéndose eco del desafío de Putnam, si la IA ya ha enseñado a los filósofos algo de importancia sobre la mente. Putnam piensa que no y sustenta su concepción con un dictamen curiosamente retórico: la IA ha fallado por completo, durante un cuarto de siglo, en la resolución de problemas que la filosofía ha fallado en resolver durante dos milenios. Tiene razón, creo, pero eso no me impresiona.[9] Es como si un filósofo llegara a la conclusión de que hay que despreciar la biología contemporánea diciendo que los biólogos no han hecho mucho para contestar a la pregunta de qué es la vida. Por cierto, no lo han hecho; han formulado mejores preguntas que deben disolver o redireccionar la curiosidad de los filósofos.

Más aún, los filósofos (entre todas las personas) deberían apreciar que las soluciones a los problemas no son el único regalo valioso; ¡problemas nuevos realmente duros son igualmente buenos! Equiparando la curiosa retórica de Putnam, yo ofrezco como la mejor contribución de la IA a la filosofía un problema epistemológico nuevo, no resuelto, ignorado por generaciones de filósofos: el problema del marco. Platón estuvo a punto de verlo. En el *Teeteto* exploró brevemente las implicaciones de una maravillosa analogía:

Sócrates: Consideremos ahora si el conocimiento es una cosa que puedes poseer sin llevarlo encima, en la misma forma en que un hombre que ha capturado algunos pájaros salvajes —palomas o lo que fuere— y los mantiene en una pajarera que ha hecho para ellos en su casa. En cierto sentido podemos decir que los «tiene» todo el tiempo, en la medida en que los posee, ¿no es así?

Teeteto: Sí.

Sócrates: Pero en otro sentido el no «tiene» ninguno, aunque ha obtenido el control sobre ellos, ahora que los ha encerrado en una prisión separada; puede tomarlos y mantenerlos cuando quiera atrapando al pájaro que elija y luego dejarlo ir; y depende de él hacerlo tan a menudo como le plazca.[10]

Platón veía que la mera posesión del conocimiento (como los pájaros en una pajarera) no es suficiente; se debe poder comandar lo que uno posee. Para hacer las cosas bien, debe poder conseguirse que la pieza correcta de conocimiento venga volando en el momento correcto (en tiempo real, como dicen los ingenieros). Pero él subestimó la dificultad de este truco y por lo tanto subestimó la clase de teoría que se debería tener de la organización del conocimiento para explicar nuestros talentos de encantadores de pájaros. Ni Platón ni ningún filósofo posterior, en la medida en que sé, vio esto como un problema de la epistemología por sí mismo, dado que las demandas de eficiencia y robustez empalidecen hasta la invisibilidad cuando se las compara con las demandas filosóficas de certidumbre; de este modo, empero, el problema surgió en las manos de la IA.[11]

Igual de importante para los filósofos como los nuevos problemas y las nuevas soluciones, sin embargo, es la nueva materia prima, y la IA la ha proporcionado en abundancia. Ha proporcionado una plétora de objetos sobre los cuales pensar, sistemas individuales con todas sus particularidades que son mucho más vívidos y temperamentales que los sistemas que yo (para nombrar a alguien) podría llegar a soñar en un experimento pensante. Esta no es una cosecha trivial. Comparemos la filosofía de la mente (el estudio analítico de los límites, oportunidades e implicaciones de las posibles teorías de la mente) con la teoría literaria de la novela (el estudio analítico de los límites, oportunidades e implicaciones de las novelas posibles). En principio se puede escribir excelente teoría literaria en ausencia de novelas como ejemplares. Aristóteles, por ejemplo, pudo en principio haber escrito un tratado anticipado sobre las fuerzas y debilidades, poderes y problemas de diversos tipos posibles de novela. Al teórico literario de hoy no se le exige que examine los ejemplares existentes, pero ellos son, para decir lo menos, una muleta útil. Extienden el rango imaginativo y la seguridad de incluso los más brillantes teóricos y proporcionan obstáculos estimulantes a las generalizaciones y conclusiones entusiastas. Las miniteorías, esbozos y modelos de la IA pueden no ser grandes novelas, pero son lo mejor que tenemos a la fecha, y así como las novelas mediocres son a menudo una recompensa para los teóricos literarios (llevan sus deficiencias en sus mangas) igualmente las malas teorías, los modelos fracasados y los presentimientos sin esperanzas en IA son una recompensa para los

filósofos de la mente. Pero usted tiene que leerlos para tener el beneficio.

Quizás el mejor ejemplo actual de este beneficio es la ola de entusiasmo hacia los modelos conexionistas. Durante años, los filósofos de la mente han estado dirigiendo sus manos, vaga y esperanzadamente, en dirección a esos modelos, completamente incapaces de concebirlos en detalle pero seguros hasta sus huesos de que eso tenía que ser posible. (Mi propio primer libro, *Content and Consciousness*, es un buen ejemplo de esa vaga forma de teorización[12]). Otros filósofos han estado igualmente seguros de que todas esas estrategias se hallaban condenadas (Jerry Fodor es un buen ejemplo). Ahora, por fin, podemos examinar una multitud de objetos en esta clase anticipada y comprobar si esas corazonadas eran correctas. En principio, sin duda, podría elaborarse todo esto sin muletas; pero en la práctica, esos desacuerdos entre filósofos tienden a degenerar en posiciones endurecidas defendidas mediantes argumentaciones cada vez más oscuras, con términos cada vez más redefinidos y con morales tendenciosas tomadas de otras residencias.

Putnam sugiere que, ya que la IA es antes que nada y esencialmente una subrama de la ingeniería, ella no puede ser filosofía. Insiste especialmente en el hecho de que tenemos que desechar sus afirmaciones en el sentido de constituir una epistemología. Encuentro llamativa esta afirmación. Seguramente Hobbes, Leibniz y Descartes estaban haciendo filosofía, incluso epistemología, cuando gesticulaban y hablaban en forma muy abstracta sobre las limitaciones de los mecanismos. Lo mismo se aplica a Kant, cuando afirmaba que estaba investigando las condiciones bajo las que la experiencia era posible. Tradicionalmente los filósofos han tratado de imaginarse la fuerza combinatoria y las limitaciones inherentes de «impresiones e ideas», de «petites perceptions», «intuiciones» y «esquemas». Los investigadores en IA han formulado preguntas parecidas sobre diversos tipos de «estructuras de datos» y «representaciones procedimentales», «marcos» y «nexos» y, sí, «esquemas», ahora un poco más rigurosamente definidos. Por lo que veo, son fundamentalmente las mismas investigaciones, pero en IA se conducen bajo restricciones adicionales (y generalmente bien motivadas) y con la ayuda de innumerables conceptos más específicos.

Putnam considera que la ingeniería y la epistemología son

incompatibles. Considero esto más bien una componenda: en la medida en que una exploración especulativa en IA es más abstracta, más idealizada y está menos restringida mecánicamente, ella es «más filosófica»; ¡pero eso no significa que sea por lo tanto de mayor interés o de mayor valor para un filósofo! Por el contrario, es probablemente porque los filósofos han sido tan filosóficos (tan abstractos, idealizados y poco restringidos por supuestos mecanísticos empíricamente plausibles) que han fracasado durante tanto tiempo en explicar la mente. La IA no ha resuelto aún ninguno de los antiguos enigmas sobre la mente, pero nos ha proporcionado nuevas formas de disciplinar y extender una imaginación filosófica que sólo ahora comenzamos a explotar.

Notas

[1] Alan Turing, «Computing Machinery and Intelligence», *Mind* 59 (236) (1950):433, reimpreso en Douglas Hofstadter y Daniel Emmett, editores, *The Mind's I* (Nueva York: Basic Books, 1981), 54-67.

[2] René Descartes, *Discourse on Method* (1637), traducción de Laurence J. LaFleur, 3ª edición (Nueva York: Bobbs-Merrill, 1960), 41-42.

[3] Donald Davidson, «Mental Events», en L. Foster y J. W. Swanson, editores, *Experience and Theory* (Amherst: University of Massachusetts Press, 1970), 79-101.

[4] François Jacob, «Evolution and Tinkering», *Science* 196 (1977):1161-66.

[5] Douglas Hofstadter, «Waking Up from the Boolean Dream, or Subcognition as Computation», capítulo 26 de *Metamagical Themas* (Nueva York: Basic Books, 1985), 631-65.

[6] John H. Holland, Keith J. Holyoak, Richard E. Nisbett y Paul R. Thagard, *Induction: Processes of Inference, Learning and Discovery* (Cambridge: MIT Press, 1986).

[7] John E. Laird, Allen Newell y Paul S. Rosenbloom, «SOAR: An Architecture for General Intelligence», *Artificial Intelligence* 33 (septiembre de 1987):1-64.

[8] Douglas Lenat, Mayank Prakash y May Shepherd, «CYC: Using Commonsense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks», *AI Magazine* 6 (4) (1986):65-85.

[9] En «Artificial Intelligence as Philosophy and as Psychology», en *Philosophical Perspectives in Artificial Intelligence*, ed. de Martin Ringle (Atlantic Highlands, N.J.: Humanities Press International, 1979), y en *Brainstorms: Philosophical Essays on Mind and Psychology* (Cambridge: MIT Press, 1978), he afirmado que la IA ha resuelto lo que he llamado el problema de Hume: el problema de romper la amenaza de la regresión infinita de homúnculos que consultan (y comprenden) representaciones internas tales como las «ideas» e «impresiones» de Hume. Sospecho que Putnam afirmará, con alguna justicia, que ha sido la computación científica en general, no la IA en particular, la que ha enseñado a la filosofía a romper esta regresión.

[10] *Plato's Theaetetus*, traducción de Francis M. Cornford (Nueva York: Macmillan, 1957), 197 C-D.

[11] Daniel C. Dennett, «Cognitive Wheels: The Frame Problem of AI», en *Minds, Machines and Evolution: Philosophical Studies*, edición de Christopher Hookway (Cambridge: Cambridge University Press, 1985), reimpreso en la nueva antología *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, edición de Zenon W. Pylyshyn (Norwood, N.J.: Ablex, 1987). En esta introducción al problema del marco, explico por qué es un problema epistemológico y por qué los filósofos no se dieron cuenta de que existía.

[12] Daniel C. Dennett, *Content and Consciousness* (Atlantic Highlands, N.J.: Humanities Press International, 1969).

14
Lógica matemática en inteligencia artificial

John McCarthy

Este artículo se refiere a los programas de computadora que representan información sobre sus dominios de problemas en lenguajes lógicos matemáticos y utilizan la inferencia lógica para decidir qué acciones son apropiadas para alcanzar sus objetivos.

La lógica matemática no es un lenguaje singular. Hay muchas clases de lógica matemática, e incluso elegir una de esas clases no implica especificar el lenguaje. El lenguaje se determina declarando qué símbolos extralógicos habrán de utilizarse y qué sentencias serán consideradas axiomas. Los símbolos extralógicos son los que tienen que ver con el tema concreto que se ha de almacenar en una base de datos de computadora: por ejemplo, información sobre objetos y sus ubicaciones y movimientos.

Cualquiera sea la elección de los símbolos, todas las clases de lógica matemática comparten dos ideas. Primero, debe estar matemáticamente definido qué hileras de símbolos se consideran fórmulas de la lógica. Segundo, debe definirse matemáticamente qué inferencias de nuevas fórmulas a partir de las viejas fórmulas están permitidas. Estas ideas permiten escribir programas de computadora que deciden qué combinaciones de símbolos son sentencias y qué inferencias se permiten en un lenguaje lógico particular.

La lógica matemática se ha convertido en una rama importante de las matemáticas, y la mayoría de los lógicos trabaja en problemas que surgen del desarrollo interno del asunto. La lógica matemática también se ha aplicado a estudiar las fundamentaciones de las matemáticas, y allí ha tenido sus mayores logros. Sus fundadores, Aristóteles, Leibniz, Boole y Frege, también deseaban

---

John C. McCarthy. Profesor de computación científica y profesor Charles M. Pigott de Ingeniería en la Universidad de Stanford.

aplicarla para hacer que los razonamientos sobre asuntos humanos fueran más rigurosos. Por cierto, Leibniz era explícito sobre su objetivo de reemplazar la argumentación mediante el cálculo. Sin embargo, expresar conocimiento y razonar sobre el mundo de sentido común en términos de lógica matemática ha entrañado dificultades que parecen exigir extensiones al concepto básico de lógica, y esas extensiones sólo están comenzando a desarrollarse.

Si una computadora ha de almacenar hechos sobre el mundo y razonar con ellos, necesita un lenguaje preciso. El programa debe basarse en una idea precisa de los razonamientos permitidos, es decir, se deben derivar nuevas fórmulas a partir de las viejas. En el comienzo fue natural intentar el uso del lenguaje lógico matemático para expresar lo que un programa inteligente de computadora «sabe» que es relevante en los problemas que queremos que resuelva y para hacer que el programa utilice la inferencia lógica para decidir qué hacer. La primera propuesta para utilizar la lógica en inteligencia artificial para expresar lo que un programa conoce y cómo debería razonar, se encuentra en un ensayo que escribí en 1960. El problema de probar fórmulas lógicas como un dominio para la IA ya se había estudiado. En ese ensayo yo decía:

> El *advice taker* es un programa propuesto para resolver problemas manipulando sentencias en lenguajes formales. La principal diferencia entre él y otros programas o programas propuestos para manipular lenguajes formales (la Máquina de Teoría Lógica de Newell, Simon y Shaw y el Programa de Geometría de Herbert Gelernter) es que en los programas anteriores el sistema formal era el asunto, pero las heurísticas se hallaban todas encarnadas en el programa. En este programa los procedimientos se describirán en la medida de lo posible en el lenguaje mismo y, en particular, las heurísticas están descritas todas del mismo modo.

La principal ventaja que esperamos tenga el *advice taker* es que su conducta no podrá demostrarse meramente haciéndole aserciones, diciéndole cosas sobre su entorno simbólico y lo que se quiere de él. Hacer estas aserciones requerirá poco conocimiento por parte del programa, si es que alguno, o del conocimiento previo del *advice taker*. Puede suponerse que el *advice taker* dispondrá de una clase bastante amplia de consecuencias lógicas inmediatas de cualquier cosa que se le diga y de su conocimiento previo. Se espera que esta propiedad tenga mucho en común con lo que nos hace describir a ciertos humanos diciendo que

tienen sentido común. Diremos en consecuencia que *un programa tiene sentido común si deduce automáticamente por sí mismo una clase suficientemente amplia de consecuencias inmediatas de cualquier cosa que se le diga y que supiera de antes.*[1]

El proyecto del *advice taker*, ambicioso en 1960, sería considerado ambicioso todavía hoy y aún está lejos de ser inmediatamente realizable. La lógica matemática está especialmente lejos del objetivo de expresar la heurística en el mismo lenguaje en el que se expresan los hechos sobre los que la heurística debe actuar. Pero los investigadores comprenden mejor las razones principales para utilizar extensivamente sentencias lógicas en IA hoy de lo que lo comprendían en 1960. Expresar información en sentencias declarativas es mucho más flexible que expresarla en segmentos de programas de computadora o en tablas. Las sentencias pueden así ser válidas en contextos mucho más amplios de lo que pueden ser útiles programas específicos. Quien suministra un dato no tiene que comprender demasiado sobre la forma en que funciona el que lo recibe, sobre el modo en que habrá de usarlo o si lo usará. El mismo hecho se puede usar para muchos propósitos; se dispondrá entonces de las consecuencias lógicas de colecciones de hechos.

Los actuales programas de computadora están más o menos cerca de este objetivo, dependiendo de la medida en que utilizan los formalismos de la lógica. Comenzaré describiendo cuatro niveles de su uso:

1. Una máquina en el nivel más bajo no utiliza sentencias lógicas. Meramente ejecuta los comandos de sus programas. Todas sus «creencias» están implícitas en sus estados. Sin embargo, a menudo es apropiado adscribir creencias y objetivos al programa. Un misil puede creer que su blanco es amigo y abandonar el objetivo de alcanzarlo. A menudo es útil decir que cierta máquina hace lo que ella piensa que logrará sus objetivos. Daniel Dennett, Allen Newell y yo hemos discutido la adscripción de cualidades mentales a las máquinas.[2] La intención de los diseñadores de la máquina y la forma en que se espera que se comporte pueden describirse más fácilmente en términos de intención que con una descripción puramente física.

La relación entre las descripciones física e intencional de una máquina es más fácil de comprender en sistemas simples que

admiten con fluidez descripciones comprensibles de ambas clases. Tomemos como ejemplo un termostato. Podemos decir que cuando él cree que la temperatura es demasiado alta, activa el sistema de enfriamiento para lograr el objetivo de obtener la temperatura correcta. Algunos filósofos fastidiosos objetan tales adscripciones. A menos que un sistema posea una mente humana completa, alegan, no se debe considerar que tiene cualidades mentales en absoluto. Esta restricción es como omitir el cero y el uno del sistema numérico sobre la base de que no se necesitan números para contar conjuntos sin elementos o con un solo elemento. Por supuesto, adscribir creencias a las máquinas (y a las personas) es más importante cuando nuestro conocimiento físico es inadecuado para explicar o predecir la conducta. Puede decirse mucho más sobre la adscripción de cualidades mentales a las máquinas, pero eso no es lo que preocupa principalmente a la IA hoy en día.

2. El siguiente nivel de lógica involucra programas de computadora que ponen sentencias en la memoria de la máquina para representar sus creencias, pero utilizan reglas diferentes de las inferencias lógicas ordinarias para alcanzar conclusiones. A menudo se obtienen nuevas sentencias a partir de las viejas mediante programas *ad hoc*. Además, las sentencias que aparecen en la memoria provienen de un subconjunto (dependiente del programa) del lenguaje lógico que se está usando. Agregar ciertas sentencias al lenguaje puede incluso desfigurar el funcionamiento del programa. La lógica se utiliza en este segundo nivel en los programas «sistemas expertos», consistentes en bases de conocimientos (p.ej. listas de síntomas de enfermedades en los sistemas expertos médicos) y máquinas de inferencia (que contienen reglas, bajo la forma de instrucciones explícitas a la máquina sobre la forma de manipular la información de las bases de conocimiento). En comparación con los lenguajes de lógica del primer orden, los lenguajes que se utilizan a este nivel son más bien inexpresivos. Por ejemplo, no pueden admitir sentencias cuantificadas (p.ej. sentencias que incluyan «para todo» o «existe»), y pueden representar reglas generales en una notación separada. A menudo, las reglas no pueden ser consecuencias de la acción de un programa; deben ser puestas por un «ingeniero del conocimiento». A menudo, las razones por las que los programas tienen esta forma es por simple ignorancia, pero la razón usual para esta restricción es la razón práctica de hacer que el

programa corra más rápido y que deduzca exactamente las clases de conclusiones que su diseñador ha anticipado. Más a menudo, las reglas son implicaciones usadas en una dirección (en otras palabras, la contrapositiva de una implicación no se utiliza). Creo que la necesidad de esa inferencia especializada será temporaria y se reducirá o se eliminará mediante formas mejoradas de controlar la inferencia (por ejemplo, permitiendo que las reglas heurísticas también se expresen como sentencias, tal como se proponía en el extracto precedente de mi ensayo de 1960.

3. El tercer nivel utiliza lógica del primer orden tanto como deducción lógica. Habitualmente las sentencias se presentan como cláusulas, y los métodos de deducción se basan en el método de resolución de J. Allen Robinson.[3] Un hecho en los datos de uno de esos programas sería:

(para todo(*x*) (si (y (inst *x* hortaliza) (color *x* púrpura))
(inst *x* berenjena)))

Traducido a un lenguaje más común, el hecho se lee: «Todas las hortalizas púrpuras son berenjenas». Su estructura es típica de las cláusulas si-entonces en las bases de datos lógicas: dada cualquier *x*, si *x* satisface las condiciones estipuladas, luego *x* asegura un cierto resultado: (para todo (*x*) (si (condiciones) entonces (resultado))).

En el ejemplo, *x* debe satisfacer dos condiciones: (inst *x* hortaliza) y (color *x* púrpura). La primera condición significa que *x* debe ser una instancia específica de la clase de las hortalizas, y la segunda significa que el color de *x* debe ser púrpura. El resultado (inst *x* berenjena), significa que *x* es una instancia de las berenjena. Armado con este hecho, el programa puede parecer listo para desarrollar esta tarea:

(inst Gertrudis hortaliza)
(color Gertrudis púrpura)
(DEMOSTRAR: (inst Gertrudis) berenjena))

Traducida, la tarea es: dado el hecho de que Gertrudis es una hortaliza púrpura, demostrar que Gertrudis es una berenjena. Pero sólo con este hecho lógico, el programa no puede hacer nada con esa

tarea. Necesita un método para razonar a partir de aserciones generales sobre x no descritas, hasta llegar a aserciones particulares sobre Gertrudis. El razonamiento de la resolución de Robinson prescribe una forma para sustituir Gertrudis por x y por lo tanto unificar las cláusulas de la base de datos con las de la tarea.

Ejemplos de esos programas que se usan comercialmente son los *shells* de sistemas expertos (ART, KEE, OPS-5), programas de computadora que crean sistemas expertos genéricos. Usted le dice al programa qué hechos desea en la base de datos; el programa convierte los hechos en aserciones lógicas y luego sigue la heurística en su propia máquina de inferencia para crear una máquina de inferencia a la medida de los hechos que usted puso en el programa.

El tercer nivel de lógica se usa menos para propósitos prácticos que el segundo nivel porque las técnicas para controlar el razonamiento se encuentran aún insuficientemente desarrolladas, y es común que un programa genere muchas conclusiones inútiles antes que alcance una solución deseada. Por cierto, experiencias poco exitosas con este método[4] han conducido a usos más estrictos de la lógica (por ejemplo, el sistema STRIPS de Richard Fikes y Nils Nilsson[5]).

En conexión con esto es importante mencionar la programación lógica, introducida por primera vez en el Microplanner[6] y desarrollada desde puntos de vista diferentes por Robert Kowalski y Alain Colmerauer en la década de 1970.[7] El Microplanner fue una colección de herramientas un tanto asistemáticas, a diferencia del Prolog, un lenguaje de computadora que se funda casi por completo en una clase de programación lógica matemáticamente tratable,[8] pero la idea es la misma. Si se utiliza una clase restringida de sentencias, las llamadas cláusulas de Horn, es posible entonces utilizar una forma restringida de deducción lógica. Esto facilita el problema del control y hace posible que el programador anticipe el curso que seguirá la deducción. El problema es que mediante cláusulas de Horn sólo se puede expresar cierta clase de hechos. Sin embargo, la posibilidad de expresarse en cláusulas de Horn es una propiedad importante de un conjunto de hechos, y la programación lógica se ha utilizado con éxito en muchas aplicaciones (aunque parece improbable que domine la programación en IA, como ciertos propulsores suyos esperan).

Aunque expresan tanto los hechos como las reglas como sentencias lógicas, los sistemas de tercer nivel siguen siendo un poco

especializados. Los axiomas con los que comienza el programa no son verdades generales sobre el mundo, sino sentencias cuyo significado y verdad se limitan al estrecho dominio en el cual el programa tiene que actuar. Por esta razón, a menudo los hechos de un programa no se pueden usar como bases de datos en otros programas.

4. El cuarto nivel es todavía un objetivo. Involucra la representación de hechos generales sobre el mundo como sentencias lógicas. Una vez puestos en una base de datos, los hechos pueden ser usados por cualquier programa. Los hechos tendrían la neutralidad de propósito característica de gran parte de la información humana. Quien suministra la información no tiene que comprender los objetivos del usuario potencial o la forma en que trabaja su mente. Las formas actuales de «enseñar» a los programas de computadora involucran una educación mediante cirugía cerebral.

Una dificultad importante es que los sistemas del cuarto nivel requieren extensiones a la lógica matemática. Una clase de extensión es el razonamiento no monotónico, propuesto por primera vez a fines de la década de 1970.[9] La lógica tradicional es monotónica en el sentido siguiente. Si se infiere una sentencia *p* de una colección *A* de sentencias, y si *B* es una colección de sentencias más abarcativa, luego se puede inferir *p* a partir de *B*. Por ejemplo, hagamos que la colección *A* sean estas sentencias: todos los solteros son no casados; Juan es un soltero. Hagamos que la colección *B* sean estas sentencias: todos los solteros son no casados; Juan no tiene novias; Juan es un soltero. De ambos conjuntos de sentencias, se puede inferir la sentencia *p*: Juan es no casado. El conjunto de sentencias *A* es un modelo del conjunto: todos los *x* son *y*. Si «*w* es *y*» es verdad en todos los modelos de este conjunto general, entonces será verdad en todos los modelos de la forma general del conjunto *B*. De este modo vemos que el carácter no monotónico de la lógica tradicional no depende de los detalles del sistema lógico, sino que es bien fundamental.

Aunque gran parte del razonamiento humano corresponde a la lógica tradicional, ciertos importantes razonamientos humanos de sentido común son no monotónicos. Alcanzamos conclusiones a partir de ciertas premisas que no alcanzaríamos si ciertas otras sentencias se incluyeran entre nuestras premisas. Por ejemplo,

sabiendo que yo poseo un automóvil, usted llega a la conclusión de que en ocasiones es apropiado pedirme que lo lleve; pero cuando usted conoce el hecho adicional de que el automóvil está siendo reparado en el taller, usted no saca más esa conclusión. Algunas personas piensan que es posible tratar de salvar la monotonicidad diciendo que lo que estaba en su mente no era una regla general sobre pedir un aventón a los propietarios de automóviles, sino una regla probabilística, algo así como «En el 70 por ciento de las veces es apropiado para usted pedirme que lo lleve si yo tengo un automóvil». Hasta ahora no se ha demostrado posible elaborar la detallada epistemología de esta estrategia, es decir, determinar exactamente qué reglas probabilísticas se deben utilizar. En vez de eso, la IA ha procurado formalizar directamente el razonamiento lógico no monotónico.

El razonamiento no monotónico formalizado se encuentra en rápido desarrollo, y se han propuesto muchas clases de sistemas. Me concentraré en una estrategia denominada «circunscripción» porque la conozco, porque ha encontrado amplia aceptación y porque es quizá la estrategia más activamente investigada en la actualidad. La idea es subrayar, de entre los modelos de la colección de sentencias que ha de presuponerse, algunos modelos «preferidos» o «estándar». Los modelos preferidos son los que satisfacen cierto principio mínimo. Lo que ha de minimizarse no se decide todavía en completa generalidad, pero muchos dominios que han sido estudiados producen teorías bastante generales utilizando minimizaciones de anormalidades o del conjunto de alguna clase de entidad. Por idea no es por completo extraña. Por ejemplo, la navaja de Occam, «No multiplicar entidades sin necesidad», es uno de esos principios mínimos.

La minimización en lógica es otro ejemplo de un área de las matemáticas que se descubre en conexión con aplicaciones más que mediante el desarrollo interno normal de las matemáticas. Por supuesto, lo inverso sucede en escala todavía mayor; muchos conceptos lógicos desarrollados por razones puramente matemáticas han resultado ser de importancia para la IA.

Como ejemplo más concreto de razonamiento no monotónico, consideremos las condiciones bajo las que se debe usar un bote para cruzar un río. Consideremos ahora qué cosas pueden andar mal con un bote. Puede tener una hendidura. Puede no tener

remos, motores o velas, dependiendo del tipo de bote que sea. Sería razonablemente conveniente listar algunas de estas cosas en un conjunto de axiomas. Sin embargo, aparte de esos obstáculos que podemos esperar listar por anticipado, el razonamiento humano admitirá que pueden surgir todavía otros, pero que no se puede pensar en todos ellos por anticipado (p.ej. una barrera en el medio del río). Se puede manejar esta dificultad utilizando la circunscripción para minimizar el conjunto de cosas que impiden que el bote cruce el río, es decir, el conjunto de obstáculos a superar. Si el razonador no conoce ninguno en un caso particular, conjeturará que se puede usar el bote, pero si conoce alguno, obtendrá un resultado distinto cuando minimice.

Esta ilustración muestra que el razonamiento no monotónico es conjetural antes que riguroso. En efecto, se ha demostrado que ciertos sistemas lógicos matemáticos no se pueden extender rigurosamente y que tienen cierta clase de completud.

Sería tan engañoso conducir una discusión de esta clase enteramente sin fórmulas, como sería discutir sin fórmulas la fundamentación de la física. Por desgracia, muchas personas son incapaces de seguir las matemáticas. De modo que discutiré una formalización de Vladimir Lifschitz de un ejemplo simple llamado «el problema del disparo de Yale».[10] Drew McDermott, que se ha desalentado respecto del uso de la lógica en IA y especialmente de los formalismos no monotónicos, desarrolló el problema como desafío.[11] Los métodos de Lifschitz trabajan bien aquí, pero pienso que requieren modificaciones.

En este problema hay originalmente un arma descargada y una persona, Fred. Luego se carga el arma. Hay una espera, y luego el arma apunta a Fred y se dispara. La conclusión deseada es que Fred muere. Informalmente, las reglas son 1) que una persona viva sigue con vida hasta que algo le sucede; 2) que el acto de cargar hace que un arma quede cargada; 3) que un arma cargada permanece cargada hasta que alguien la descarga; 4) que un disparo descarga un arma; y 5) que disparando un arma cargada contra una persona mata a la persona. Pretendemos razonar como sigue: Fred permanecerá vivo hasta que se dispare un arma porque no puede inferirse nada que haga que mientras tanto le suceda algo; el arma permanecerá cargada hasta que se dispare porque no puede inferirse nada que haga que mientras tanto algo le suceda; Fred morirá cuando el arma se dispare. La parte no

monotónica del razonamiento consiste en minimizar las cosas que suceden o suponer que nada sucede sin una razón.

Las sentencias lógicas pretenden expresar esas cinco premisas, pero no dicen explícitamente que no sucedan otros fenómenos. Por ejemplo, no hay ninguna aserción en el sentido de que Fred no esté utilizando un chaleco a prueba de balas, ni se menciona alguna propiedad de los chalecos a prueba de balas. Sin embargo, una persona llegará a la conclusión de que al menos que un aspecto no mencionado de la situación se presente para impedir la muerte de Fred, él morirá. La dificultad radica en que las sentencias admiten un «modelo mínimo no intencional», para utilizar la terminología de la lógica matemática. Específicamente puede suceder que por alguna razón no especificada el arma se descargue durante la espera, de modo que Fred siga vivo. La forma en que se usaron formalismos no monotónicos (p.ej. la circunscripción y la lógica por defecto de R. A. Reiter) para formular el problema, minimizando la «anormalidad», resulta en dos posibilidades, no una. La posibilidad no intencional es que el arma se descargue misteriosamente.

Parece probable que la introducción del razonamiento no monotónico no será la única modificación de la lógica requerida para otorgar a las máquinas la capacidad para el razonamiento de sentido común. Para hacer programas que razonen sobre sus propios razonamientos y creencias (es decir, programas que posean una conciencia, aunque más no sea rudimentaria), es necesario formalizar muchas nociones intencionales (p.ej., conocimiento y creencia). Se pueden formalizar algunos de ellos en lógica del primer orden introduciendo proposiciones y conceptos como individuos.[12] Complicando estos esfuerzos están las paradojas descubiertas por Richard Montague.[13] Para evitarlas, será necesario debilitar adecuadamente los axiomas, pero aún no se ha encontrado una buena forma de hacerlo. También parece necesario formalizar la noción de contexto, pero esto se encuentra en un estado de investigación muy preliminar.[14]

IA y filosofía

La inteligencia artificial no puede evitar la filosofía. Si un programa de computadora se ha de comportar inteligentemente en el mundo real, se le debe proporcionar alguna clase de marco de referencia en el cual encajar los hechos particulares que se le

dicen o que descubre. Esto involucra al menos un fragmento de alguna clase de filosofía, aunque sea ingenua. Aquí estoy de acuerdo con lo filósofos que abogan por el estudio de la filosofía y alegan que alguien que intente ignorarla meramente se está condenando a una filosofía ingenua.

Dado que aún está muy por detrás de la performance intelectual de la gente filosóficamente ingenua, la IA probablemente se las arregle con una filosofía ingenua durante un largo tiempo. Por desgracia, no ha sido posible decir qué es una filosofía ingenua, y los filósofos ofrecen escasa orientación.

La siguiente alternativa plausible sería construir programas que busquen y representen el conocimiento de acuerdo con las convicciones de alguna filosofía que hayan propuesto los filósofos. Esto tampoco ha sido posible. O bien nadie en IA (incluyendo los filósofos retirados) entiende las teorías filosóficas lo suficientemente bien como para programar computadoras de acuerdo con sus dogmas, o bien los filósofos no han estado ni remotamente cerca de la precisión requerida. En realidad, algunas filosofías empiristas parecen ser lo suficientemente precisas, pero resultan inadecuadas cuando se pretende usarlas en el más modesto programa de computadora. En consecuencia, nosotros, los investigadores en IA, nos tenemos que arreglar solos cuando se trata de proporcionar una estructura intelectual básica a un programa. Estas son algunas cosas que pensamos que esto requeriría:

Ontología. Adopto la idea de Willard Quine de que nuestra ontología se define por el rango de variables ligadas.[15] Con esta idea, necesitamos especificar las clases de entidades que han de presuponerse, es decir, sobre qué tratan las ideas del robot. Aunque el nominalismo sugeriría, además, que las variables sólo toman como valores objetos materiales, esta teoría pronto se muestra inadecuada porque, por ejemplo, no permite que el diseñador del robot le informe sobre las cualidades de los objetos que se preservan cuando tienen lugar determinados sucesos.

Quine nos dice que «en la ciencia no hay lugar para las ideas», y argumenta en favor de esta concepción con ejemplos acerca de la dificultad de definir qué significa que dos personas tengan la misma idea.[16] Sin embargo, si un programa ha de buscar una buena idea generando montones de ideas y luego poniéndolas a prueba,

necesita algún criterio para decir cuándo ha probado ya cierta idea. De esta manera, parecerían ser necesarias ideas como objetos, pero no se ha descubierto aún cómo evitar la dificultad que menciona Quine. Los sistemas actuales de IA no pueden enumerar ideas.

Libre albedrío. Los robots que planeamos construir son sistemas completamente deterministas. Sin embargo, un robot sofisticado debe decidir qué hacer considerando las diversas cosas que puede hacer y escogiendo la que tenga las mejores consecuencias a la vista de los objetivos que se le han fijado. Para hacerlo, debe ser capaz de representar «Yo puedo hacer *A* y puedo hacer *B*, pero *B* parece mejor, de modo que aunque puedo hacer *A*, no lo haré». ¿Qué significa para un robot creer «Yo puedo, pero no lo haré»? Es un sistema determinístico, de modo que hará *A* o no lo hará. Patrick J. Hayes y yo hemos ofrecido algunas propuestas para resolver el problema del libre albedrío en los robots.[17]

Razonamiento no monotónico. Los programas de IA requieren formas de saltar a conclusiones sobre la base de evidencia insuficiente.

Los intentos de los investigadores de IA de determinar un marco de referencia intelectual lo suficientemente preciso para programar sistemas de IA ya han conducido a ciertas concepciones filosóficas, tanto para tomar partido en algunas antiguas controversias filosóficas como para formular propuestas que consideramos nuevas. Discutiré dos tópicos:

1. *Incrementalismo o modestia.* Los hechos sobre los efectos de las acciones y de otros sucesos que se han colocado en las bases de datos de los programas de IA no son demasiado generales. No son ni siquiera tan generales como lo que un cuestionario elicitaría entre gente ingenua, no digamos lo suficientemente generales como para satisfacer a la gente familiarizada con la literatura filosófica. Sin embargo, en ciertos casos alcanza para determinar la acción apropiada para alcanzar una meta. Observar las limitaciones en estos casos lleva a un avance adicional. Esta es una metodología útil incluso cuando los objetivos son filosóficos. Se pueden diseñar formalismos que pueden usarse en sistemas activos y mejorarlos cuando sus defectos resulten evidentes.

El filósofo puede alegar que los sistemas en actividad son demasiado triviales para resultarle interesantes. Podría estar equi-

vocado, porque es evidente que las investigaciones filosóficas de la acción no han advertido fenómenos importantes que surgen tan pronto como uno trata de diseñar un sistema para planificar la acción. Aquí van dos ejemplos. Primero, las ideas sobre la asociación, que se remontan por lo menos a Mill, pasando por los conductistas, son demasiado vagas como para ser programadas en absoluto. Segundo, los filósofos no han advertido la mayor parte del carácter no monotónico del razonamiento involucrado en la toma cotidiana de decisiones. Para la IA es importante no sólo que el investigador sea capaz de desarrollar sus ideas, sino también que el programa sea capaz de mejorar incrementalmente su comportamiento, ya sea aceptando consejos del usuario o aprendiendo con la experiencia, y esas mejoras requieren nuevos lenguajes para expresar el conocimiento. Por ejemplo, un bebé considera inicialmente que la palabra madre es un nombre propio, luego un nombre general para las mujeres adultas, y todavía más tarde la designación de una relación. Pienso que antes que podamos tener programas de computadora con la inteligencia general y la flexibilidad lingüística de un niño, los investigadores de IA tienen que desarrollar lenguajes con «tolerancia a la elaboración». Por ejemplo, un lenguaje semejante debería permitir que el uso de la palabra madre se desarrolle como se describió arriba sin perder la información más vieja. La tolerancia a la elaboración es un tópico de investigación actual en IA.

2. *Objetividad.* A pesar de la concepción de la realidad última que tengamos, al diseñar robots necesitamos hacer que la concepción del mundo del robot sea una realidad externa sobre la cual él tiene y puede obtener sólo un conocimiento parcial. No tendremos éxito si diseñamos el robot para que considere el mundo como una mera estructura construida en su información sensorial. Aquí tiene que haber una teoría (podría llamársela una meta-epistemología) que relacione la estructura de un mundo, un buscador de conocimiento en ese mundo, el canal de interacción entre el buscador de conocimiento y el resto del mundo, las reglas del buscador de conocimiento para decidir qué aserciones sobre el mundo son significativas y las reglas del buscador de conocimiento para aceptar evidencia sobre el mundo y lo que el buscador de conocimiento puede descubrir. Si las reglas son demasiado restrictivas (como quizá lo son en algunas filosofías operacionalistas de la ciencia), el buscador de conocimiento, considerando las aserciones insuficientemente operacionales

como para ser significativas, será incapaz de descubrir hechos básicos con respecto al mundo.

Observaciones

Gran parte de lo que pretendo decir involucra afirmar una postura sobre cuestiones que son polémicas incluso dentro de la IA.

Creo, por ejemplo, que la inteligencia artificial se comprende mejor como una rama de la computación científica que como una rama de la filosofía. La IA tiene que ver con métodos para alcanzar objetivos en situaciones en que la información disponible tiene cierto carácter complejo. Los métodos que acostumbramos usar se relacionan con el problema que presenta la situación y son similares independientemente de que quien resuelve el problema sea un ser humano, un marciano o un programa de computadora.

Inicialmente, muchos eran demasiado optimistas sobre el tiempo que insumiría alcanzar una inteligencia de nivel humano. El optimismo era natural porque sólo se habían identificado unas pocas dificultades. Ahora ya se han identificado suficientes dificultades como para establecer la IA como una de las ciencias más difíciles. Alcanzar una inteligencia de nivel humano puede insumir cinco años, o puede insumir quinientos.

Todavía no está claro cómo caracterizar las situaciones que requieren inteligencia. Evidentemente, ellas son situaciones abiertas. Incluso en un juego como el ajedrez, donde las reglas son fijas, los métodos para decidir una movida son de carácter abierto: todo el tiempo se inventan nuevas formas de pensar las posiciones del ajedrez.

La IA ya ha identificado ciertos métodos de reconocimiento de patrones, de búsquedas heurísticas en árboles de posibilidades y de representación de información mediante reglas y aprendizaje. Todavía hay que caracterizar otros métodos, especialmente métodos para representar problemas como colecciones de subproblemas que se pueden examinar por separado para obtener resultados que se puedan usar en el estudio de sus interacciones.

Acercarse a la IA a través de la lógica no es la única estrategia que puede resultar exitosa. Por ejemplo, con el tiempo es posible que resulten exitosas estrategias ligadas más estrechamente a la biología, aun cuando la mayoría de las estrategias biológicamente motivadas que se han ensayado desde los 50 se han secado.

Hay mucha controversia en torno de las implicaciones de la IA para la filosofía, un tema sobre el que hay posturas muy fuertes. La IA tiende a dar sustento a concepciones racionalistas y realistas de los problemas filosóficos, antes que a concepciones empiristas, fenomenológicas o idealistas. Alienta una estrategia fragmentadora a la filosofía de la mente, en la que las cualidades mentales se consideran por separado, antes que como partes de un gran paquete. Esto es así porque algunos sistemas tienen cualidades mentales importantes, pero más bien limitadas.

Hay muchos problemas en la formalización del sentido común, y muchas estrategias para resolverlos aguardan ser exploradas. Dos mil años de filosofía sólo poseen a este respecto una relevancia limitada. En mi opinión, la discusión adecuada de estos problemas es inevitablemente técnica en su mayor parte, habida cuenta de los formalismos lógicos concretos que se usan. El cálculo de situación que se utiliza posee limitaciones importantes bien conocidas. El formalismo *resulta (e,s)*, utilizado en IA para expresar las consecuencias de la acción y otros sucesos, tiene que modificarse para manejar el tiempo continuo. Se necesita un formalismo muy diferente para expresar hechos sobre sucesos concurrentes. El «cálculo de sucesos» de Robert Kowalski y Mark Sergot es candidato a reunir ambos requerimientos.[18]

El estudio de la IA puede conducir a una metaepistemología matemática análoga a la metamatemática: un estudio de la relación entre las reglas del conocedor para aceptar la evidencia y un mundo en el que el conocedor está inserto. Este estudio puede derivar en teoremas matemáticos sobre la posibilidad de que ciertas estrategias intelectuales lleven al descubrimiento de ciertos hechos sobre el mundo. Pienso que con el tiempo esta posibilidad revolucionará la filosofía.[19]

Notas

[1] John McCarthy, «Programs with Common Sense», en *Proceedings of the Teddington Conference on the Mechanization of Thought Processes* (Londres: Her Majesty's Stationery Office, 1960), 77-84.

[2] Daniel C. Dennett, «Intentional Systems», *Journal of Philosophy* 68 (4) (25 de febrero de 1971):25; Allen Newell, «The Knowledge Level», *AI Magazine* 2 (2) (1981):87-106; y John McCarthy, «Ascribing Mental Qualities to Machines», en

Philosophical Perspectives in Artificial Intelligence, edición de Martin Ringle (Brighton, Sussex: Harvester Press, 1979), 1-20.

[3] J. Allen Robinson, «A Machine-oriented Logic Based on the Resolution Principle», *Journal of the Association for Computing Machinery* 12 (1) (1965):23-41.

[4] Cordell Green, «Application of Theorem Proving to Problem Solving», *International Joint Conference on Artificial Intelligence* 1 (1969):219-39.

[5] Richard Fikes y Nils Nilsson, «STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving», *Artificial Intelligence* 2 (3, 4) (enero de 1971): 189-208.

[6] Gerald J. Sussman, Terry Winograd y Eugenie Charniak, «Micro-planner Reference Manual», Report AIM-203A (Cambridge: Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1971).

[7] Robert Kowalski, *Logic for Problem Solving* (Amsterdam: North-Holland, 1979); la primera implementación de Prolog fue desarrollada por Alain Colmerauer, de la Universidad de Marsella en 1971, pero sólo fue descrita en los documentos internos del grupo.

[8] Un texto reciente sobre programación lógica es el de Leon Sterling y Ehud Shapiro, *The Art of Prolog* (Cambridge: MIT Press, 1986).

[9] John McCarthy, «Epistemological Problems of Artificial Intelligence», en *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (Cambridge: Massachusetts Institute of Technology, 1977); McCarthy, «Circumscription - A Form of Non-Monotonic Reasoning», *Artificial Intelligence* 13 (1,2) (1980):27-39; McCarthy, «Applications of Circumscription to Formalizing Common Sense Knowledge», *Artificial Intelligence* 28 (1) (1986):89-116. Véase también Raymond A. Reiter, «A Logic for Default Reasoning», *Artificial Intelligence* 13 (1,2) (1980):81-132; y Drew McDermott y Jon Doyle, «Non-Monotonic Logic I», *Artificial Intelligence* 13 (1) (1980):41-72.

[10] Vladimir Lifschitz, «Formal Theories of Action», reporte preliminar en el volumen 2 de *Proceedings of the International Joint Conference on Artificial Intelligence* (Los Altos, Calif.: Morgan-Kaufmann, 1977), 966-72.

[11] Drew McDermott, «A Critique of Pure Reason», *Computational Intelligence*, 1988.

[12] John McCarthy, «First Order Theories of Individual Concepts and Propositions», en *Machine Intelligence* 9, ed. de Donald Michie (Edinburgo: University Of Edinburg Press, 1979), 129-48.

[13] Richard Montague, «Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability», *Acta Philosophica Fennica* 16 (1963):153-67, reimpreso en Richard Montague, *Formal Philosophy* (New Haven: Yale University Press, 1974).

[14] Michael Genesereth y Nils Nilsson han escrito el mejor texto general sobre la aproximación lógica a la IA, *The Logical Foundations of Artificial Intelligence* (Los Altos, Calif.: Morgan-Kaufmann, 1987).

[15] Willard V. Quine, *Quiddities* (Cambridge: Harvard University Press, 1987).

[16] Ibid.

[17] John McCarthy y Patrick J. Hayes, «Some Philosophical Problems from the Standpoint of Artificial Intelligence», en *Machine Intelligence* 4, ed. Donald Michie (Nueva York: American Elsevier, 1969).

[18] Robert Kowalski y Sergot Marek, *A Logic-Based Calculus of Events* (Londres: Department of Computing, Imperial College, 1985).

[19] Una versión más matemática de este ensayo se publicó como «Artificial Intelligence, Logic and Formalizing Common Sense», en Richmond H. Thomason (ed.), (Norwell, Mass.: Kluwer Academic Publishers, 1989).

NOTA FINAL

Le recordamos que este libro ha sido prestado gratuitamente para uso exclusivamente educacional bajo condición de ser destruido una vez leído. Si es así, destrúyalo en forma inmediata.

Súmese como voluntario o donante, para promover el crecimiento y la difusión de la Biblioteca





Ciencias cognitivas
Serie CLA·DE·MA



gedisa
editorial

Stephen R. Graubard *(comp.)*

El nuevo debate sobre la inteligencia artificial

El nuevo debate sobre la inteligencia artificial no confronta ya a los que creen y a los que no creen en sus posibilidades, sino a filósofos, informáticos y neurobiólogos que conocen en profundidad el alcance de la inteligencia artificial y que evalúan sus verdaderas perspectivas en comparación con la actividad mental humana.

Mientras que Hubert L. Dreyfus sugiere encontrar los orígenes de la discusión metafóricamente en la oposición entre Heidegger y Husserl y Hilary Putnam invita –con Chomsky– a ver la inteligencia artificial en sus límites técnicos, David L. Walz nos presenta "máquinas verdaderamente inteligentes" y J. D. Cowan y D. E. Sharp muestran los sorprendentes paralelismos estructurales entre las redes neuronales y los circuitos de la inteligencia artificial.

En opinión de John McCarthy, "la inteligencia artificial no puede evitar la filosofía. Si un programa de computadora ha de comportarse inteligentemente en el mundo real, se le debe proporcionar un marco de referencia en el cual encajar los hechos particulares que se le dicen o que descubre."

Más allá de posiciones a favor y en contra, los diseños de texturas cerebrales mismos que sólo poco a poco se van descubriendo, parecen superar en su comportamiento emergente y holístico todo lo que hasta ahora se ha entendido por "explicación científica". Lejos de estar cerrado, el debate sobre la inteligencia artificial de hecho sólo está en sus inicios y no podemos prever quién nos deparará mayores sorpresas: la mente humana o la inteligencia artificial.



ISBN 84-7432-466-1
9 788474 324662
302376